mathematics

MDPI

Article

Mathematical Framework for Mixed Reservation- and Priority-Based Traffic Coexistence in 5G NR Systems

Daria Ivanova [1,*], Yves Adou [1], Ekaterina Markova [1], Yuliya Gaidamaka [1,2] and Konstantin Samouylov [1]

[1] Applied Informatics and Probability Department, Peoples' Friendship University of Russia (RUDN University), 6 Miklukho-Maklaya St, 117198 Moscow, Russia
[2] Institute of Informatics Problems, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences (FRC CSC RAS), 44-2 Vavilov St, 119333 Moscow, Russia
* Correspondence: ivanova-dv@rudn.ru

Abstract: Fifth-generation (5G) New Radio (NR) systems are expected to support multiple traffic classes including enhanced mobile broadband (eMBB), ultra-reliable low-latency communications (URLLC), and massive machine-type communications (mMTC) at the same air interface. This functionality is assumed to be implemented by utilizing the network slicing concept. According to the 3rd Generation Partnership Project (3GPP), the efficient support of this feature requires statistical multiplexing and, at the same time, traffic isolation between slices. In this paper, we formulate and solve a mathematical model for a class of Radio Access Network (RAN) slicing algorithms that simultaneously include resource reservation and a priority-based service discipline allowing us to incur fine granularity in the service processes of different traffic aggregates. The system is based on a queueing model and allows parametrization by accounting for the specifics of wireless channel impairments. As metrics of interest, we utilize K-class session drop probability, K-class session pre-emption probability, and system resource utilization. To showcase the capabilities of the model, we also compare performance guarantees provided for URLLC, eMBB, and mMTC traffic when multiplexed over the same NR radio interface. Our results demonstrate that the performance trade-off is dictated by the offered traffic load of the highest priority sessions: (i) when it is small, mixed reservation/priority scheme outperforms the full reservation mechanism; (ii) for overloaded conditions, full reservations provides better traffic isolation. The mixed strategy is beneficial to traffic aggregates with short-lived lightweight sessions, such as URLLC and mMTC, while the reservation only scheme works better for elastic eMBB traffic. The most important feature is that the mixed strategy allows resource utilization to be improved up to 95%, which is 10–15% higher compared to the reservation-only scheme while still providing isolation between traffic types.

Keywords: 5G NR; network slicing; radio access network; mathematical modeling; queueing theory; pre-emptive priority; resource reservation

MSC: 60K25; 60K30; 90B18; 90B22

Citation: Ivanova, D.; Adou, Y.; Markova, E.; Gaidamaka, Y.; Samouylov, K. Mathematical Framework for Mixed Reservation- and Priority-Based Traffic Coexistence in 5G NR Systems. *Mathematics* **2023**, *11*, 1046. https://doi.org/10.3390/math11041046

Academic Editor: Ivan Ganchev

Received: 11 January 2023
Revised: 14 February 2023
Accepted: 16 February 2023
Published: 18 February 2023

1. Introduction

Network slicing is an essential functionality of future fifth-generation (5G) New Radio (NR) cellular systems [1,2]. It can allow resources to be allocated and isolated for different traffic classes or mobile virtual network operators (MVNOs) on the same physical network infrastructure, e.g., mobile physical network operators (MPNOs) [3]. In both cases, traffic aggregates may have different quality of service (QoS) requirements in terms of throughput, latency, and drop probability as stated in the International Telecommunications Union Radiocommunication Sector (ITU-R) M.2410 [4].

QoS provisioning to traffic slices needs to be provided in an end-to-end way [5,6]. The critical part is the Radio Access Network (RAN), where time-varying radio channel

conditions may deteriorate statically provisioned performance guarantees [7,8]. In addition, according to the unified opinion of the 3rd Generation Partnership Project (3GPP), ITU-R, and the Groupe Speciale Mobile Association (GSMA) [9,10], the utilized slicing scheme should satisfy two contradictory requirements—traffic isolation and efficient usage of radio resources. The former requirement can be implemented by utilizing static or dynamic resource allocation policies [11,12], while the latter can be implemented via traffic prioritization schemes [13,14]. The joint dynamics of schemes accounting for both types of requirements is rather complex. To this end, mathematical models are needed to understand the trade-offs between user-centric key performance indicators (KPI), such as session loss and pre-emption probabilities, and system-centric ones, such as the efficiency of system resource utilization.

1.1. Related Work

There have been a number of studies addressing the question of network slicing at the air interface. Specifically, Ref. [15] focuses on two key features of slicing: traffic isolation and automated management. In doing so, a pre-emption-based prioritization (PP) scheme, "merging" the resource allocation and traffic prioritization schemes, is proposed. To evaluate or estimate the so-called PP scheme, a queuing system model analyzing the functioning of a single base station (BS) accommodating multiple services with different QoS requirements is given. Concretely, this paper considers each service-oriented slice to be assigned an overall share of radio resources, including guaranteed ones utilized by neighboring slices. As one key result, the proposed PP scheme can achieve 100% gain in terms of blocking probabilities with respect to a predefined baseline. In addition, Ref. [11] investigates the following features of slicing: flexible priority-based traffic isolation, fair QoS-aware resource allocation, and efficient usage of radio resources. The authors proposed a slicing scheme bolstering the traffic isolation and maintaining the efficient usage of radio resources, and represented it by utilizing a queuing model with three 5G services having uniform data rate requirements at one BS. This slicing scheme takes advantage of the complete partitioning and complete sharing policies' key features. In practice, this paper considers the users of each service-oriented slice to be ensured a minimum data rate, with the possibility to achieve higher data rates whenever radio resources are free. The proposed slicing scheme can achieve 90% gain in terms of average user satisfaction index, and reduce the session drop probability by an order of magnitude dependent on the baseline.

Authors in [13] studied the industrial deployment of 5G with simultaneous support of enhanced mobile broadband (eMBB) and ultra-reliable low-latency communication (URLLC) services. The effectiveness of coexistence strategies is achieved via explicit prioritization. In addition, the authors consider service strategies in which URLLC traffic can be offloaded onto device-to-device (D2D) connections with and without explicit reservation of a fraction of resources for direct connections. As a result, the authors concluded that a D2D-aware strategy, in which the BS explicitly reserves resources for direct connection, is significantly superior to strategies in which no explicit reservation is used, as well as the strategy without support for D2D connections.

Recently, studies suggesting the use of machine learning (ML) for network slicing have started to appear. Specifically, Ref. [16] utilizes deep reinforcement learning (DRL) technology and proposes a hybrid hard–soft slicing framework, guaranteeing service level agreements (SLAs) and maximizing the spectrum efficiency given some isolation constraints. Technically, the paper considers the users of each service-oriented slice to be able to utilize the radio resources of a newly configured service-oriented slice neighboring existing ones. As the main result, the SLAs can be guaranteed all the time with the proposed hybrid slicing able to achieve near-optimal performance in terms of SLA satisfaction ratio, isolation degree, and spectrum efficiency. As the main drawback, the proposed slicing cannot satisfy mixed SLAs such as latency and reliability. The authors in [17] investigate the impact of traffic in one slice on QoS parameters experienced by the traffic in another slice. They develop a data-driven slicing and allocation model by using ML algorithms,

where resources between network slices are intelligently redistributed in accordance with prescribed QoS parameters. The study in [18] considers an experimental 5G network prototype with the ability to configure radio resources for network slices using ML solutions based on real-time performance metrics. The obtained results confirm that the ML-based approaches outperform the traditional ones and improve the utilization of resources while guaranteeing the QoS parameters.

1.2. Contributions

In this paper, by utilizing the tools of the queuing theory, we formulate and solve a general slicing RAN problem for K traffic aggregates utilizing both resource reservation and pre-emptive-priority service discipline. In the considered system, each slice is assigned a certain dedicated share of radio resources g_k such that $\sum_k g_k < C$, where C is the overall amount of radio resources. The shared pool of resources, $C - \sum_k g_k$ is regulated by the pre-emptive priority service procedure to improve the degree of statistical multiplexing. The system is then solved for user- and system-centric KPI, session drop/pre-emption probability as well as system resource utilization, by utilizing the queuing-theoretic formalism. The proposed system allows to satisfy inherently contradictory requirements of having a strong degree of isolation between traffic classes and high efficiency of resource usage. The performance of the proposed system is demonstrated by utilizing a three-class slicing scheme serving an ultra-reliable low-latency service, an enhanced mobile broadband service, and a massive machine-type communication (mMTC) service at the same radio interface.

The main contributions of our study are the following:

- Mathematical framework for mixed reservation- and priority-based network slicing at the radio interface along with parameterization by accounting for wireless channel specifics that include numerous special traffic isolation strategies such as full reservation and priority;
- Observation that the mixed reservation- and priority-based strategy allows extremely high resource utilization to be maintained, approaching 95% in overloaded system conditions, while still providing a strong degree of traffic isolation;
- Observations that short-lived lightweight low-priority traffic, such as mMTC, is best suited for the mixed reservation–priority strategy while elastic eMBB traffic benefits more from the full reservation strategy in terms of both drop and pre-emption probabilities.

The rest of the paper is organized as follows. First, in Section 2 we introduce our system model. The mathematical model is formalized and solved for performance metrics of interest in Section 3. Numerical results of the coexistence of three services, URLLC, eMBB, and mMTC, are presented in Section 4. Conclusions are drawn in the last section.

2. System Model

In this section, we introduce the considered system model. We start by describing the scenario, then proceed by specifying the details of the radio part, and finally, introduce the metrics of interest.

2.1. Considered Scenario

We consider a single base station (BS) deployment with a circular coverage area of radius r, see Figure 1a. We assume that network slicing is utilized to deliver K services having different QoS requirements. Further, for brevity, we use "k-type session" to mean "user's request for service of type k", $k = 1, \ldots, K$. The geometric locations of user equipment (UE)-generating sessions are assumed to be uniformly distributed in the cell coverage. The height of BS and UE is assumed to be h_{BS} and h_{UE}, respectively.

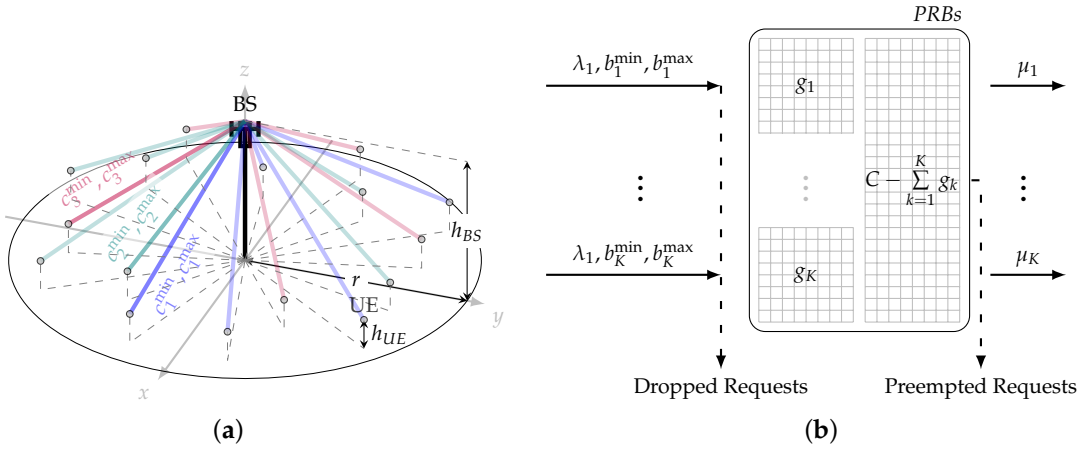


Figure 1. The considered system model and its formalization in terms of a queueing system: (**a**) system model, (**b**) queueing model.

The considered system is formalized in terms of queueing theory, see Figure 1b. We consider a queueing system with K types of traffic assumed to be partially elastic with the minimum and the maximum resource requirements, $1 \leq b_k^{\min} \leq b_k^{\max}, k = 1, \ldots, K$, expressed in primary resource blocks (PRB). The resource requirements correspond to the rates $c_k^{\min}$ and $c_k^{\max}$ and are derived below by utilizing the radio part models. We also denote $\mathbf{b}_{\min} = (b_1^{\min}, \ldots, b_K^{\min})^T$, $\mathbf{b}_{\max} = (b_1^{\max}, \ldots, b_K^{\max})^T$. Requests for service of type k arrive according to the homogeneous Poisson process with intensity λ_k. The session service time is exponentially distributed with the mean μ_k^{-1}. We assume that $g_k, k = 1, \ldots, K$, PRBs from C PRBs are reserved for each type of service requests, $\sum_{k=1}^{K} g_k < C$. Then $c = C - \sum_{k=1}^{K} g_k$ represents the shared pool of resources available for any type of traffic.

2.2. Radio Part

The impact of the radio part manifests itself in two critical parameters: (i) coverage of the BS r and (ii) resource request characterization, $b_k^{\min}$, $b_k^{\max}$. These two quantities can be obtained by utilizing session rate requirements $c_k^{\min}$, $c_k^{\max}$ and radio part models, including propagation, antenna, and blockage models specified in [19] depending on the considered deployment and operational frequency. Below, we illustrate the basic steps involved in parameterization [20,21]. The exposure provided below is applicable for both microwave (μWave) and millimeter wave (mmWave) 5G NR systems, see, for example, [22].

The coverage radius r of the cell is assumed to be the maximum separation between the UE and the NR BS, such that the UE is not in an outage. The signal-to-interference-plus-noise ratio (SINR) at the 3D distance y is provided by

$$S(y) = \frac{P_B G_B G_U M_{SF} L(y)}{N_0 B + M_I}, \tag{1}$$

where P_B is the NR BS emitted power; G_B and G_U are the BS and UE antenna gains, respectively, obtained from [23]; N_0 is the thermal noise power spectral density; B is the system bandwidth; $M_{SF} \sim N(0, \sigma_{SF})$ is the shadowing; M_I is the cross-cell interference margin; and $L(y)$ is the path loss that can be obtained from TR 38.901 [19].

$$L_{dB}(y) = \beta + 10\zeta \log_{10} y + 20 \log_{10} f_c, \tag{2}$$

where f_c is the carrier frequency measured in GHz, β, and ζ are the parameters provided in [19] that depend on the utilized band and propagation conditions.

To derive r, we assume the worst-case scenario for millimeter wave communications, such that UE at the cell edge is in non-line-of-sight (nLoS) conditions and is also blocked.

Observe that the path loss in (2) can be represented in the linear scale by utilizing the model in the form $Ay^{\zeta_{nLoS}}$, where $A = 10^{2\log_{10} f_c + \beta}$. The 3D distance y is given by

$$y = \sqrt{r^2 + [h_{BS} - h_{UE}]^2}. \tag{3}$$

By solving (1) and (3) with respect to r, we have

$$r = \sqrt{\left(\frac{P_B G_B G_U M_{SF}}{10^{\frac{\beta}{10}} f_c^2 (N_0 B + M_I) S_{th}}\right)^{\frac{2}{\zeta_{nLoS}}} - (h_{BS} - h_{UE})^2}, \tag{4}$$

where $\zeta_{nLoS} = 3.19$, $\beta = 52.4$ dB, $S_{th} = -8.97$ dB is the SINR threshold corresponding to the minimum feasible modulation and coding scheme (MCS) in 5G NR [19,24].

Select r such that the UE at the cell edge spends no more than 5% of time in outage conditions. The corresponding value of M_{SF} is provided by solving

$$M_{SF} = \sqrt{2}\text{erfc}^{-1}(2p_{out})\sigma_{\text{SF}}, \tag{5}$$

where $p_{out} = 0.05$, $\text{erfc}^{-1}(\cdot)$ is the inverse complementary error function, σ_{SF} is the standard deviation of the shadow fading distribution in the nLoS state tabulated in [19].

Once radius r is obtained, one may characterize the required resources $b_k^{\min}$, $b_k^{\max}$ $k = 1, 2, \ldots, K$, to satisfy session rate requirements, $c_k^{\min}$, $c_k^{\max}$. Recalling that UEs are assumed to be uniformly distributed in the coverage of NR BS with the probability density function (pdf) in the form $f(y) = 2y/r^2$, $0 < y < r$ [25], the mean spectral efficiency can be obtained as follows

$$E[C_e] = \int_0^r \frac{2y}{r^2} \log_2[1 + S(y)]dy, \tag{6}$$

where $S(y)$ is the SINR at 3D distance y.

Accounting for the rate of applications, $c_k^{\min}$, $c_k^{\max}$, $k = 1, 2, \ldots, K$, and available bandwidth at NR BS, B, one may now utilize the mean spectral efficiency to estimate the mean amount of resources requested by considered UEs

$$b_k^{\max} = \frac{c_k^{\max}}{E[C_e]}, \; b_k^{\min} = \frac{c_k^{\min}}{E[C_e]}, \tag{7}$$

and then, by utilizing the employed NR numerology (1/2 for μWave and 3/4 for mmWave), the mean amount of requested resources is further converted into PRBs [24].

Note that the 5G NR system considered in this paper, in addition to the sub-6 GHz band, may also operate in the mmWave band. At the same time, the model developed further is general and does not depend on the type of propagation environment and operational frequency. To this aim, here we provided parameterization for sub-6 GHz systems that captures the specifics of a wireless channel. However, the specifics of mmWave, namely, the dependence on the line-of-sight (LoS) propagation of signals, can be taken into account by using the approaches described, for example, in our previous work [13] for industrial deployment, or in the studies of other authors [20,22] for general deployments.

2.3. Metrics of Interest

In this paper, we are interested in two types of metrics for the proposed slicing scheme: (i) user-centric and (ii) system-centric. The first types of metrics include session drop and session pre-emption probabilities as a function of the number of traffic classes, resource reservations, and priority order. The main system-centric metric of interest is the efficiency of resource utilization.

3. Mathematical Model

The behavior of the queueing system specified in Section 2 can be described by the K-dimensional continuous-time Markov chain (CTMC) $\{(N_1(t),\dots,N_K(t)), t \geq 0\}$, where $N_k(t)$ captures the number of k-type requests in the system at time t. Denote $N_k = \lfloor (C - \sum_{i\neq k} g_i)/b_k^{\min} \rfloor$ as the maximum number of k-type sessions in service, then $\mathbf{n} = (n_1,\dots,n_K)$, $n_k = 0,\dots,N_k$, is the number of k-type sessions that are currently in the system, $i,k = 1,\dots,K$. Additionally, let us denote $N_k^g = \lfloor g_k/b_k^{\min} \rfloor, k = 1,\dots,K$ as the maximum guaranteed number of k-type sessions in service. The considered process is defined over the following state space

$$\mathcal{X} = \{\mathbf{n} : 0 \leq n_k \leq N_k, k = 1,\dots,K, \sum_{i=1}^{K} \max\{n_i b_i^{\min}, g_i\} \leq C\}. \tag{8}$$

Due to the partially elastic nature of the considered sessions, the amount of resources, $b_k(\mathbf{n})$, $b_k^{\min} \leq b_k(\mathbf{n}) \leq b_k^{\max}$, available to the k-type sessions is equally distributed between them and depends on the state $\mathbf{n} \in \mathcal{X}$. That is, we have

$$b_k(\mathbf{n}) = \\ = \min\left\{ \frac{C - \max\{\sum_{i\neq k} g_i, \sum_{i=1}^{k-1} \max\{n_i b_i(\mathbf{n}), g_i\} + \sum_{i=k+1}^{K} \max\{n_i b_i^{\min}, g_i\}\}}{n_k}, b_k^{\max} \right\}. \tag{9}$$

3.1. Model without Pre-Emption

Let us start with a partial sharing strategy with a non-pre-emptive priority. In the state space $\mathcal{X}$ we can identify three important sets, namely $\mathcal{S}_k$, $\mathcal{B}_k$, and $\mathcal{S}_k^{\max}$, $k = 1,\dots,K$. The former set, $\mathcal{S}_k$, $k = 1,\dots,K$, called the "accepting" set, contains all the states in which sessions are accepted to the system and is provided by

$$\mathcal{S}_k = \{\mathbf{n} \in \mathcal{X} : n_k < N_k, \sum_{i=1}^{k-1} \max(n_i b_i(\mathbf{n} + \mathbf{e}_k), g_i) + \\ + (n_k + 1) b_k^{\min} + \sum_{i=k+1}^{K} \max(n_i b_i^{\min}, g_i) \leq C\}, k = 1,\dots,K. \tag{10}$$

The second set of states $\mathcal{B}_k$, $k = 1,\dots,K$, called the "loss" set, is a set of system states in which arriving sessions are dropped,

$$\mathcal{B}_k = \{\mathbf{n} \in \mathcal{X} : n_k = N_k \vee \sum_{i=1}^{k-1} \max(n_i b_i(\mathbf{n} + \mathbf{e}_k), g_i) + \\ + (n_k + 1) b_k^{\min} + \sum_{i=k+1}^{K} \max(n_i b_i^{\min}, g_i) > C\}, k = 1,\dots,K. \tag{11}$$

Finally, in the accepting set $\mathcal{S}_k$, we can select a subset $\mathcal{S}_k^{\max}$, $k = 1,\dots,K$, a set of system states in which the arriving sessions will be accepted to the system using $b_k^{\max}$ PRBs, i.e.,

$$\mathcal{S}_k^{\max} = \{\mathbf{n} \in \mathcal{X} : n_k < \left\lfloor \frac{C - \sum_{i\neq k} g_i}{b_k^{\max}} \right\rfloor, \sum_{i=1}^{k-1} \max(n_i b_i(\mathbf{n} + \mathbf{e}_k), g_i) + \\ + (n_k + 1) b_k^{\max} + \sum_{i=k+1}^{K} \max(n_i b_i^{\min}, g_i) \leq C\}, k, i = 1,\dots,K. \tag{12}$$

Consider now the process of session admission to the system. Particularly, when a new session arrives to the system, the following may happen:

- If (i) upon arrival of the new session of type k there are more than $b_k^{\min}, k = 1, \ldots, K$, PRBs available, and (ii) the current amount of sessions in service is smaller than $N_k, k = 1, \ldots, K$, this session is accepted at the system;
- In any other case, the session is rejected.

Denote by $\mathbf{p} = \{p(n_1, \ldots, n_K), (n_1, \ldots, n_K) \in \mathcal{X}\}$ the steady-state probability distribution of the CTMC $\mathbf{X}(t)$. Since the considered stochastic process $\mathbf{X}(t)$ is reversible, by solving the system of local balance equations,

$$\begin{aligned} \lambda_k p(\mathbf{n}) &= (n_k + 1)\mu_k p(\mathbf{n} + \mathbf{e}_k), \\ p(\mathbf{n} + \mathbf{e}_k) &= \frac{\lambda_k}{(n_k + 1)\mu_k} p(\mathbf{n}), \end{aligned} \tag{13}$$

recursively, we obtain the steady-state probability distribution of the system in the product-form (14), that is,

$$\begin{aligned} p(n_1, \ldots, n_K) &= \frac{\rho_1^{n_1}}{n_1!} \cdot \ldots \cdot \frac{\rho_K^{n_K}}{n_K!} p(0, \ldots, 0),\ n_k = 0, 1, \ldots, N_k, \\ p(0, \ldots, 0) &= \left(\sum_{n_1=0}^{N_1} \ldots \sum_{n_K=0}^{N_K} \frac{\rho_1^{n_1}}{n_1!} \cdot \ldots \cdot \frac{\rho_K^{n_K}}{n_K!} \right)^{-1}, \end{aligned} \tag{14}$$

where $\rho_k = \lambda_k / \mu_k, k = 1, \ldots, K$.

Once the steady-state distribution vector $\mathbf{p}$ is found, we may proceed to determine the performance measures of the system:

- Drop probability of k-type session, p_{B_k}, is given by

$$p_{B_k} = \sum_{n_1=0}^{N_1} \ldots \sum_{n_K=0}^{N_K} p(\mathbf{n}) I\{\mathbf{n} \in \mathcal{B}_k\}, k = 1, \ldots, K; \tag{15}$$

- Average amount of resources occupied by i-type sessions, $\bar{k}_i$, is given by

$$\bar{k}_i = \sum_{n_1=0}^{N_1} \ldots \sum_{n_K=0}^{N_K} n_i b_i(\mathbf{n}) p(\mathbf{n}) I\{\mathbf{n} \in \mathcal{X}\}, i = 1, \ldots, K; \tag{16}$$

- The fraction of utilized resources, U, is provided as

$$U = \sum_{i=1}^{K} \sum_{n_1=0}^{N_1} \ldots \sum_{n_K=0}^{N_K} n_i b_i(\mathbf{n}) p(\mathbf{n}) I\{\mathbf{n} \in \mathcal{X}\}. \tag{17}$$

3.2. Model with Pre-Emption

Consider now the system with the pre-emptive priority service. The behavior of this queuing system is similar to the previous one in the sense that the considered process is defined over the same state space $\mathcal{X}$ as in (8). We assume that the services have different priorities: the highest priority is for the first type of services, the lowest is for K-type of services. Priority service is implemented in such a way that in the case of insufficient resources in the system to provide the k-type service with a minimum requirement, the service of one or more lower priority sessions in the shared pool of resources c could be pre-empted.

Figure 2 illustrates the algorithm for selecting sessions that could be pre-empted when a new session arrives to the system. We assume that $\mathbf{m} = (m_1, \ldots, m_K)$—the numbers of sessions that could be pre-empted. Given the initial condition $i = K, i > j, m_k = 0$, $k = 1, \ldots, K$, consider the steps of the algorithm.

Step 1. Check if there are enough resources for arriving at j-type session. If the condition is met, the session is accepted. If the condition is not met, go to Step 2.

Step 2. Check if the service of the i-type session in the shared pool c can be pre-empted. If the condition is met, increase m_i by one and go to Step 1. If the condition is not met, go to Step 3.

Step 3. Check if there are sessions in the system with a higher priority $i-1 > j$. If the condition is met, go to Step 2 with $i = i-1$. If the condition is not met, the session is rejected.

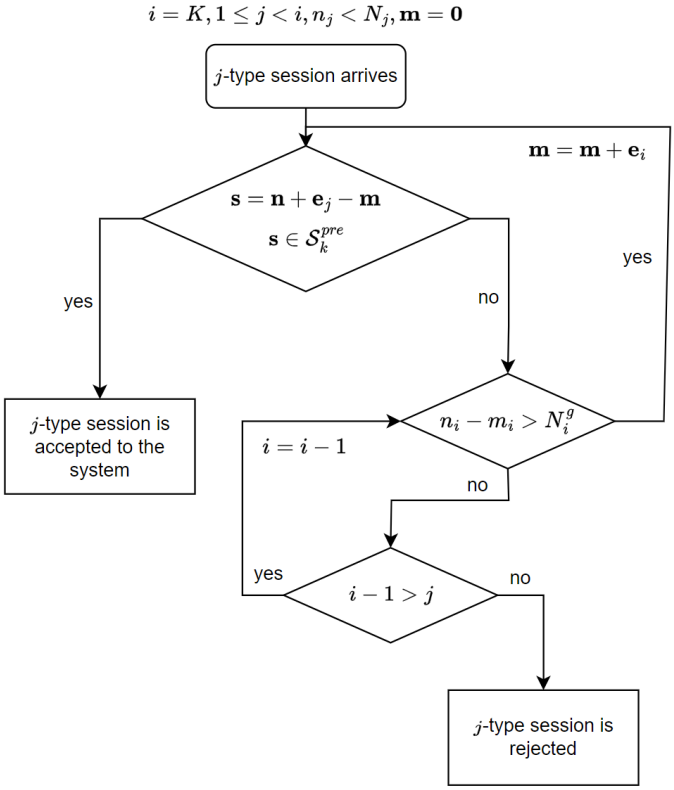


Figure 2. Algorithm for selecting sessions that could be pre-empted.

Similarly to the previous model, we also define accepting, loss, and accepting with the maximum number of PRB sets of system states, as $\mathcal{S}_k^{pre}$, $\mathcal{B}_k^{pre}$, and $\mathcal{S}_k^{\max}$, $k = 1, \ldots, K$, respectively. The first set, $\mathcal{S}_k^{pre}$, is defined as

$$\mathcal{S}_k^{pre} = \{\mathbf{n} \in \mathcal{X} : n_k < N_k, (\sum_{i=1}^{k-1} \max(n_i b_i(\mathbf{n}+\mathbf{e}_k), g_i) + \sum_{i=k+1}^{K} \max(n_i b_i^{\min}, g_i) +$$
$$+ (n_k+1)b_k^{\min} \le C \vee (k < K, \sum_{i=1}^{k-1} \max(n_i b_i(\mathbf{n}+\mathbf{e}_k), g_i) + \sum_{i=k+1}^{K} \max(n_i b_i^{\min}, g_i) +$$
$$+ (n_k+1)b_k^{\min} - \sum_{i=k+1}^{K} b_i^{\min}(n_i - N_i^g) \cdot I(n_i > N_i^g) \le C))\}, k = 1, \ldots, K, \quad (18)$$

the second set, $\mathcal{B}_k^{pre}$, is provided by

$$\mathcal{B}_k^{pre} = \{\mathbf{n} \in \mathcal{X} : n_k = N_k \vee \sum_{i=1}^{k-1} \max(n_i b_i(\mathbf{n}+\mathbf{e}_k), g_i) + \sum_{i=k+1}^{K} \max(n_i b_i^{\min}, g_i) +$$
$$+ (n_k+1)b_k^{\min} - \sum_{i=k+1}^{K} b_i^{\min}(n_i - N_i^g) \cdot I(n_i > N_i^g) > C\}, k = 1, \ldots, K, \quad (19)$$

and the last set, $\mathcal{S}_k^{\max}$, is defined in (12).

For pre-emptive service, the system of these sets should also be complemented with the pre-emption set of states, Π_k, $k = 1, \ldots, K-1$, where a k-type session is accepted to the system, causing the pre-emption of lower priority sessions. This set is given by

$$\Pi_k = \{\mathbf{n} \in \mathcal{X} : n_k < N_k, \sum_{i=1}^{k-1} \max(n_i b_i(\mathbf{n}+\mathbf{e}_k), g_i) + \sum_{i=k+1}^{K} \max(n_i b_i^{\min}, g_i) + \\ + (n_k+1)b_k^{\min} > C, \sum_{i=1}^{k-1} \max(n_i b_i(\mathbf{n}+\mathbf{e}_k), g_i) + \sum_{i=k+1}^{K} \max(n_i b_i^{\min}, g_i) + \\ + (n_k+1)b_k^{\min} - \sum_{i=k+1}^{K} b_i^{\min}(n_i - N_i^g) \cdot I(n_i > N_i^g) \le C\}, k = 1, \ldots, K-1. \quad (20)$$

Consider now the process of session admission at the system. Particularly, when a new session arrives to the system, the following may happen:

- If (i) upon arrival of the new k-type session there are more than $b_k^{\min}, k = 1, \ldots, K$, PRBs available, and (ii) the current amount of k-type sessions in service is smaller than $N_k, k = 1, \ldots, K$, this session is accepted to the system;
- If (i) the arriving k-type session, $k = 1, \ldots, K-1$, observes less than $b_k^{\min}$ free PRBs in the system, and (ii) the current amount of k-type sessions in service is smaller than N_k, and (iii) there are more than $b_k^{\min}$ PRBs occupied by lower priority sessions in the shared pool of resources, c, the session is admitted to the system causing the pre-emption of $\sum_{i=2}^{K} m_i$ lower priority sessions;
- In any other case, the session is rejected.

By utilizing the rules specified above, we can fully characterize the stochastic process $\mathbf{X}(t)$ describing the service of URLLC and eMBB sessions with the partial reservation strategy and pre-emptive priority. The arbitrarily chosen "central" state of the process is shown in Figure 3.

$$a(\mathbf{n}, \mathbf{n}') = \begin{cases} \lambda_k, & \text{if } \mathbf{n}' = \mathbf{n} + \mathbf{e}_k, \mathbf{n} \in \mathcal{S}_k, k = 1, \ldots, K, \\ & \text{or } \mathbf{n}' = \mathbf{n} - \mathbf{m} + \mathbf{e}_k, \mathbf{n} \notin \mathcal{S}_k, \mathbf{n} \in \mathcal{S}_k^{pre}, n_i > m_i, k, i = 1, \ldots, K, \\ n_k\mu_k, & \text{if } \mathbf{n}' = \mathbf{n} - \mathbf{e}_k, n_k > 0, k = 1, \ldots, K; \\ *, & \text{if } \mathbf{n}' = \mathbf{n}, k = 1, \ldots, K; \\ 0, & \text{otherwise.} \end{cases} \quad (21)$$

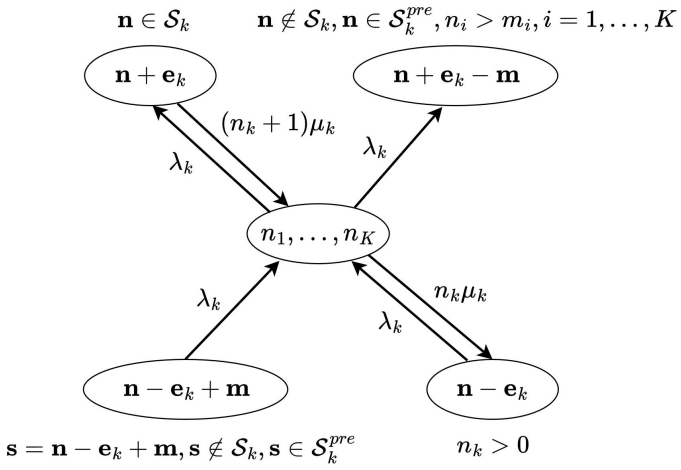


Figure 3. Transition probabilities of the central state.

As opposed to the non-pre-emptive system, the CTMC $\mathbf{X}(t)$ is non-reversible, implying that no product-form solution is available for this system. In this case, the steady-state distribution can be estimated numerically by solving the following set of linear equations:

$$\mathbf{p}^T\mathbf{A} = \mathbf{0}^T, \mathbf{p}^T\mathbf{1} = 1, \tag{22}$$

where $\mathbf{A}$ is an infinitesimal generator having elements $a(\mathbf{n}, \mathbf{n'})$ defined in (21), with the shorthand notation $*$ provided by

$$* = -[\sum_{k=1}^{K} \lambda_k I\{\mathbf{n} \in \mathcal{S}_k\} + \lambda_k I\{\mathbf{n} \notin \mathcal{S}_k, \mathbf{n} \in \mathcal{S}_k^{pre}, n_i > m_i, i = 1, \ldots, K\} + n_k\mu_k]. \tag{23}$$

Once the steady-state distribution vector $\mathbf{p}$ is obtained, one may proceed with the performance measures of the considered system that can be expressed as follows:

- Drop probability of k-type session, $p_{B_k^{pre}}$, is given by

$$p_{B_k^{pre}} = \sum_{n_1=0}^{N_1} \ldots \sum_{n_K=0}^{N_K} p(\mathbf{n}) I\{\mathbf{n} \in \mathcal{B}_k^{pre}\}; \tag{24}$$

- The pre-emption probability, p_{pre_k}, that is, the probability that arbitrarily chosen sessions are dropped during ongoing service when k-type session is accepted to the system

$$p_{pre_k} = \sum_{n_1=0}^{N_1} \ldots \sum_{n_K=0}^{N_K} p(\mathbf{n}) I\{\mathbf{n} \in \Pi_k\}. \tag{25}$$

Note that $\bar{k}_i$, U are defined in (16) and (17), respectively.

4. Numerical Results

In this section, we proceed to provide an illustrative example of the proposed mixed reservation- and priority-based traffic coexistence strategy. Specifically, we consider three services that have to be supported by the 5G NR air interface—URLLC, eMBB, and mMTC. According to the 5G International Mobile Telecommunications-2020 (IMT-2020) requirements specified in [4] and network slicing recommendations in 3GPP Technical Specification (TS) 23.501 and TS 38.300, URLLC should receive the highest priority and be served as there are absolutely no other types of traffic in the system reaching the drop probability of 10^{-5}. On the other hand, mMTC requirements are the loosest out of all these three types of services, requiring that no more than 1% of traffic is not delivered within 10 s. The eMBB service is characterized by the balanced requirements. Thus, in our numerical example, we assign the priorities accordingly.

The default system parameters utilized to produce the reported results in this section are shown in Table 1. Note that here, we explicitly compare two slice isolation schemes: (i) full reservation, where the whole set of resources, $C = 39$ PRBs, is equally divided between the slices, and (ii) partial reservation with priorities, where a set of resources is allocated to the shared pool. Note that in both cases, URLLC traffic is well isolated from the rest of the traffic and, thus, we will only consider performance degradation experienced by eMBB and mMTC.

Table 1. Parameters utilized for numerical assessment.

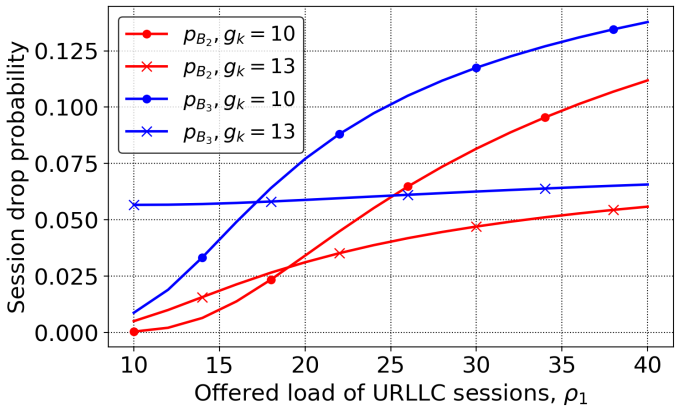
Parameter	Value
Arrival intensity of URLLC sessions	10,000 sess./s
Arrival intensity of eMBB sessions	1 sess./s
Arrival intensity of mMTC sessions	10,000 sess./s
Mean service time of URLLC sessions	1 ms
Mean service time of eMBB sessions	10 s
Mean service time of mMTC sessions	1 ms
Overall amount of resources in the system	39 PRB
Amount of reserved resources for URLLC	10, 13 PRB
Amount of reserved resources for eMBB	10, 13 PRB
Amount of reserved resources for mMTC	10, 13 PRB
Resource requirement of URLLC sessions	1 PRB
Minimum resource requirement of eMBB sessions	1 PRB
Maximum resource requirement of eMBB sessions	3 PRB
Resource requirement of mMTC sessions	1 PRB

We start with Figure 4 with the two critical metrics of interest, eMBB and mMTC session drop, p_{B_2}, p_{B_3}, and pre-emption probabilities, p_{pre_2}, p_{pre_3}, for two considered schemes as a function of the offered traffic load of URLLC, $\rho_1 = \lambda_1 b_1^{\min}/\mu_1$. Note that the full reservation scheme is indicated by $g_k = 13$ PRBs, while the mixed reservation- and priority-based is indicated by $g_k = 10$. In the latter case, there is the shared pool of nine PRBs that are shared between all types of traffic. By analyzing the session drop probability demonstrated in Figure 4a, one may observe that the mixed reservation- and priority-based mechanism outperforms full reservation in terms of the mMTC drop probability for low and moderate values of the ρ_1. For both eMBB and mMTC traffic, the system is characterized by two regimes: (i) up until approximately $\rho_1 = 15$, the mixed scheme performs better, (ii) after $\rho_1 = 15$, the full reservation scheme outperforms the mixed one. The rationale is that prior to $\rho_1 = 15$ both eMBB and mMTC traffic efficiently utilize the shared pool of resources available. However, when the offered traffic load of highest priority URLLC sessions further increases, they start to occupy the shared pool of resources, aggressively leading to the performance loss of eMBB and mMTC traffic.

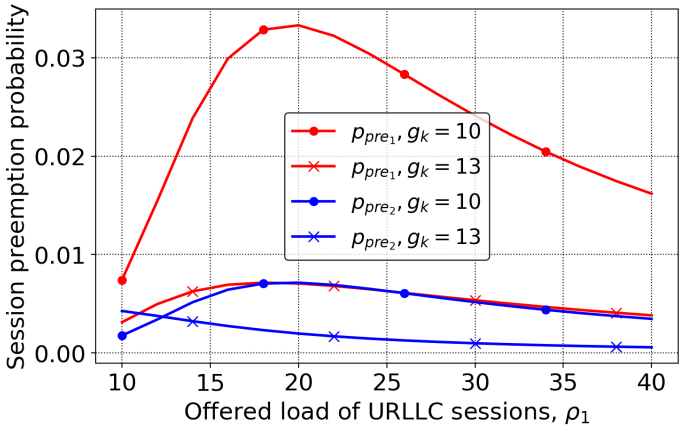
The price one has to pay for better session drop probability in the mixed reservation- and priority-based scheme is higher session pre-emption probability demonstrated in Figure 4b for the same system parameters. Here, we see that the session pre-emption probability is lower for the full reservation scheme across almost the whole range of the URLLC-offered traffic load ρ_1. Notably, the mMTC traffic suffers less with the corresponding probabilities being quite close to each other. The rationale here is that short-lived lightweight traffic is better suited for priority service due to fine granularity in terms of the number of sessions that need to be interrupted to accommodate the arriving higher priority session. However, the elastic eMBB traffic suffers most with the difference reaching three times at approximately $\rho_1 = 20$. Note that the pre-emption of a rare but long eMBB session has much higher negative impact compared to mMTC sessions, thus making the full reservation scheme better suited for the former type of traffic.

We finally proceed with the resource utilization of the system demonstrated in Figure 5 for the same set of input parameters as in Figure 4, as a function of URLLC offered traffic load, ρ_1, and eMBB maximum requested rate, $b_2^{\max}$. By analyzing the results presented in Figure 5a, we observe that across the whole range of ρ_1 the mixed strategy shows consistently better results by approximately 10%. Notably, in overloaded conditions with a certain level of isolation supported, the resource utilization is just 5% off from 100%. In fact, this is the best side of the considered mixed strategy. Qualitatively similar results are also observed in Figure 5b, where resource utilization of the system as a function of the eMBB maximum requested rate $b_2^{\max}$ is shown. Here, we might observe that as the system starts to be overloaded, the mixed scheme allows 95% of utilization to be reached, while the

full reservation schemes remain at approximately 80%, resulting in slightly higher gains of around 15%.

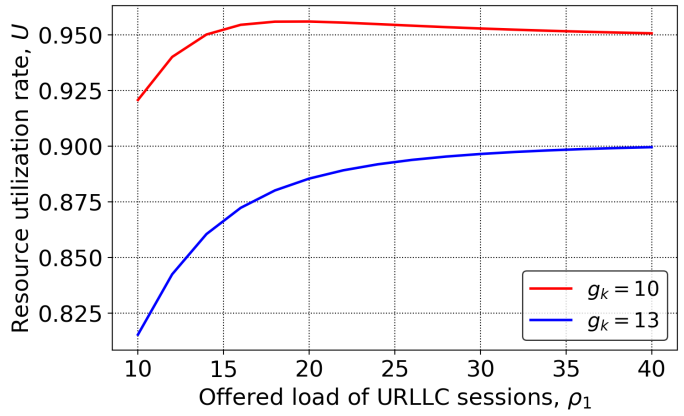


(**a**)



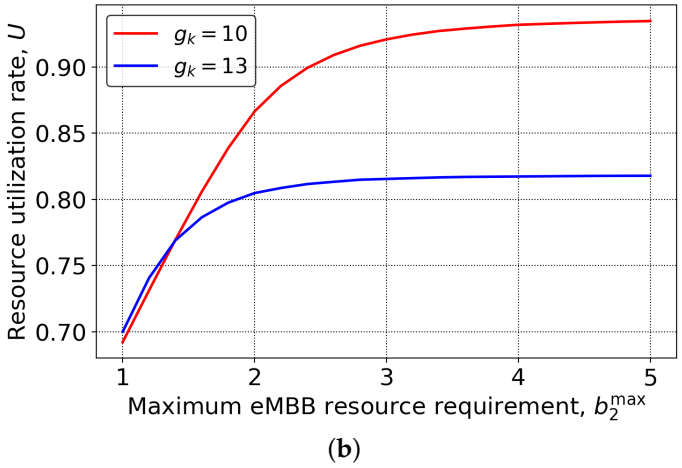
(**b**)

Figure 4. Drop and pre-emption probabilities: (**a**) eMBB and mMTC drop probabilities, (**b**) URLLC and eMBB pre-emption probabilities.



(**a**)

Figure 5. *Cont.*



(**b**)

Figure 5. System resource utilization: (**a**) varying ρ_1, (**b**) varying $b_2^{\max}$.

5. Conclusions

In this paper, motivated by the need for simple mathematical models for RAT network slicing, we proposed and solved the model for mixed reservation- and priority-based traffic coexistence. The proposed framework aims to provide the balance between traffic isolation and efficient use of resources and also includes numerous service strategies as its special cases including full reservation, priority-only strategies, as well as mixed strategy without pre-emption. The system is formulated and solved by utilizing the tools of the queuing theory. We also demonstrated how the input parameters can be related to the wireless channel specifics including antennas and propagation models.

Our numerical results demonstrate that the proposed mixed reservation- and priority-based strategy allows resource utilization to be improved up to 95% while still providing isolation between traffic types in highly overloaded conditions. Compared to the full reservation strategy, the gains are in the range 10–15% across a wide range of system parameters. The mixed scheme is also better in terms of session drop probabilities for low and moderate offered traffic load of the highest priority sessions. However, elastic eMBB traffic benefits more from the full reservation strategy in terms of both drop and pre-emption probabilities.

One of the important advantages of the proposed approach is the simplicity of implementation, since for one of the models the steady-state probability distribution can be obtained by solving the system of local balance equations in a product form. However, for the model with pre-emption, the product-form solution is not available, so the steady-state distribution can only be estimated numerically by solving a system of linear equations. In addition, we considered 5G NR, which may operate in a mmWave or sub-6 GHz band. The proposed model is general and does not depend on the type of propagation environment and operational frequency.

In our future studies, the performance of the proposed algorithm on specific traffic sets and per-slice QoS/QoE (quality of Eexperience) will be evaluated. Moreover, propagation models that are closer to reality will also be utilized.

Author Contributions: Conceptualization, E.M., Y.G., and K.S.; methodology, D.I., E.M. and Y.G.; software, D.I.; validation, E.M. and Y.G.; formal analysis, E.M. and D.I; investigation, D.I.; resources, E.M. and D.I.; data curation, E.M.; writing—original draft preparation, D.I. and E.M.; writing—review and editing, Y.A., E.M. and Y.G.; visualization, Y.A.; supervision, E.M. and Y.G.; project administration, E.M., Y.G. and K.S.; funding acquisition, E.M. and K.S. All authors have read and agreed to the published version of the manuscript.

Funding: Sections 2–4 were written by Daria Ivanova, Yves Adou, and Yuliya Gaidamaka under the support of the Russian Science Foundation, project no. 22-79-10053, https://rscf.ru/en/project/22-79-10053/, accessed 1 September 2022. This paper has been supported by the RUDN University Strategic Academic Leadership Program (recipients Ekaterina Markova, Konstantin Samouylov, Sections 1 and 5).

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. ITU-T. *Requirements of the IMT-2020 Network*; ITU-T Rec. Y.3101; ITU-T: Geneva, Switzerland, 2018.
2. ETSI. *System Architecture for the 5G System*; 3GPP TS 23.501 Version 15.2.0 Release 15; ETSI: Sophia Antipolis, France, 2018.
3. Li, Y.; Zheng, J.; Li, Z.; Liu, Y.; Qian, F.; Bai, S.; Liu, Y.; Xin, X. Understanding the ecosystem and addressing the fundamental concerns of commercial MVNO. *IEEE/ACM Trans. Netw.* **2020**, *28*, 1364–1377. [CrossRef]
4. Series, M. *Minimum Requirements Related to Technical Performance for IMT-2020 Radio Interface(s)*; ITU-R M.2410-0; ITU: Geneva, Switzerland, 2017.
5. Nakao, A.; Du, P.; Kiriha, Y.; Granelli, F.; Gebremariam, A.A.; Taleb, T.; Bagaa, M. End-to-end network slicing for 5G mobile networks. *J. Inf. Process.* **2017**, *25*, 153–163. [CrossRef]
6. Li, X.; Ni, R.; Chen, J.; Lyu, Y.; Rong, Z.; Du, R. End-to-end network slicing in radio access network, transport network and core network domains. *IEEE Access* **2020**, *8*, 29525–29537. [CrossRef]
7. ETSI. *UMTS QoS Concept and Architecture*; 3GPP TS 23.107 Version 4.1.0 Release 4; ETSI: Sophia Antipolis, France, 2001.
8. Yan, M.; Feng, G.; Zhou, J.; Sun, Y.; Liang, Y.C. Intelligent resource scheduling for 5G radio access network slicing. *IEEE Trans. Veh. Technol.* **2019**, *68*, 7691–7703. [CrossRef]
9. GSM Association. *GSM Association Official Document NG.116 (11/2020)*; Generic Network Slice Template; Version 4.0; GSM Association: London, UK, 2020.
10. ITU-T. *Framework for the Support of Network Slicing in the IMT-2020 Network*; ITU-T Rec.Y.3112; ITU-T: Geneva, Switzerland, 2018.
11. Yarkina, N.; Correia, L.M.; Moltchanov, D.; Gaidamaka, Y.; Samouylov, K. Multi-tenant resource sharing with equitable-priority-based performance isolation of slices for 5G cellular systems. *Comput. Commun.* **2022**, *188*, 39–51. [CrossRef]
12. Naddeh, N.; Jemaa, S.B.; Eddine Elayoubi, S.; Chahed, T. Proactive RAN Resource Reservation for URLLC Vehicular Slice. In Proceedings of the 2021 IEEE 93rd Vehicular Technology Conference (VTC2021-Spring), Helsinki, Finland, 25–28 April 2021; pp. 1–5. [CrossRef]
13. Ivanova, D.; Markova, E.; Moltchanov, D.; Pirmagomedov, R.; Koucheryavy, Y.; Samouylov, K. Performance of Priority-Based Traffic Coexistence Strategies in 5G mmWave Industrial Deployments. *IEEE Access* **2022**, *10*, 9241–9256. [CrossRef]
14. Yang, W.; Li, C.P.; Fakoorian, A.; Hosseini, K.; Chen, W. Dynamic URLLC and eMBB Multiplexing Design in 5G New Radio. In Proceedings of the 2020 IEEE 17th Annual Consumer Communications Networking Conference (CCNC), Las Vegas, NV, USA, 10–13 January 2020; pp. 1–5. [CrossRef]
15. Adou, Y.; Markova, E.; Gaidamaka, Y. Modeling and Analyzing Preemption-Based Service Prioritization in 5G Networks Slicing Framework. *Future Internet* **2022**, *14*, 299. [CrossRef]
16. Zhang, H.; Pan, G.; Xu, S.; Zhang, S.; Jiang, Z. A Hard and Soft Hybrid Slicing Framework for Service Level Agreement Guarantee via Deep Reinforcement Learning. In Proceedings of the 2022 IEEE 95th Vehicular Technology Conference: (VTC2022-Spring), Helsinki, Finland, 19–22 June 2022; IEEE: Piscataway, NJ, USA, 2022. [CrossRef]
17. Kumar, N.; Ahmad, A. Machine learning-based QoS and traffic-aware prediction-assisted dynamic network slicing. *Int. J. Commun. Netw. Distrib. Syst.* **2022**, *28*, 27–42. [CrossRef]
18. Salhab, N.; Langar, R.; Rahim, R.; Cherrier, S.; Outtagarts, A. Autonomous Network Slicing Prototype Using Machine-Learning-Based Forecasting for Radio Resources. *IEEE Commun. Mag.* **2021**, *59*, 73–79. [CrossRef]
19. 3GPP. *Study on Channel Model for Frequencies from 0.5 to 100 GHz (Release 14)*; 3GPP TR 38.901 V14.1.1; 3GPP: Sophia-Antipolis, France, 2017.
20. Kovalchukov, R.; Moltchanov, D.; Gaidamaka, Y.; Bobrikova, E. An accurate approximation of resource request distributions in millimeter wave 3GPP new radio systems. In Proceedings of the Internet of Things, Smart Spaces, and Next Generation Networks and Systems: 19th International Conference, NEW2AN 2019, and 12th Conference, ruSMART 2019, Proceedings 19, St. Petersburg, Russia, 26–28 August 2019; Springer: Berlin/Heidelberg, Germany, 2019; pp. 572–585.
21. Sopin, E.; Moltchanov, D.; Daraseliya, A.; Koucheryavy, Y.; Gaidamaka, Y. User Association and Multi-connectivity Strategies in Joint Terahertz and Millimeter Wave 6G Systems. *IEEE Trans. Veh. Technol.* **2022**, *71*, 12765–12781. [CrossRef]
22. Moltchanov, D.; Sopin, E.; Begishev, V.; Samuylov, A.; Koucheryavy, Y.; Samouylov, K. A tutorial on mathematical modeling of 5G/6G millimeter wave and terahertz cellular systems. *IEEE Commun. Surv. Tutor.* **2022**, *24*, 1072–1116. [CrossRef]
23. Constantine, A.B. Antenna theory: Analysis and design. In *Microstrip Antennas*, 3rd ed.; John Wiley & Sons: Hoboken, NJ, USA, 2005.

24. 3GPP. *NR; Physical Channels and Modulation (Release 15)*; 3GPP TR 38.211; 3GPP: Sophia-Antipolis, France, 2017.
25. Moltchanov, D. Distance distributions in random networks. *Ad. Hoc. Netw.* **2012**, *10*, 1146–1166. [CrossRef]