To the analysis of the dynamic assignment of radio resources in wireless networks with a network slicing mechanism

Ekaterina V. Bobrikova[a], Anna A. Platonova[a], Sergey Ya. Shorgin[b] and Yuliya V. Gaidamaka[a,b]

[a]*Peoples' Friendship University of Russia (RUDN University), 6, Miklukho-Maklaya St., Moscow, 117198, Russia*
[b]*Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences (FRC CSC RAS), 44-2, Vavilov St., Moscow, 119333, Russia*

Abstract

Network Slicing is one of the latest technologies of modern telecommunication systems. Network Slicing involves dividing the 5G physical architecture into multiple virtual networks or slices. Each slice has its own characteristics and is aimed at solving a particular business problem. In the nearest future it is expected that Network Slicing principle will radically change the approach, in particular, of mobile operators to support vertical applications with specific and rigorous performance requirements. Network resource tenants can manage these requirements. The static assignment of resources to tenants, that is, the assignment of resources on a permanent basis, is a sufficient condition for fulfilling terms of Service Level Agreement (SLA), but this assignment can lead to the significant inefficiencies of the frequency resource and to the high cost of renting it for virtual mobile operators. As an alternative solution, the method of a dynamic resource sharing can be proposed. In this paper we consider the principle of setting network slices using an utility function. This principle implements resource planning mechanisms for tenants taking into account traffic requirements. These mechanisms allow to differentiate slices and prioritize services that correspond to slices.

Keywords

5G, elastic traffic, Network Slicing, virtualization, resource allocation, scheduling, infrastructure sharing

1. Introduction

It is expected that in the coming years, mobile Internet traffic will not only continue to grow rapidly, but will also change and reorient in connection with an unprecedented number and variety of network-connected devices and the necessity to support a wide range of new applications. This will lead to a significant increase in the costs of network operators: costs that are not comparable with the growth of operators' profits.

Workshop on information technology and scientific computing in the framework of the X International Conference Information and Telecommunication Technologies and Mathematical Modeling of High-Tech Systems (ITTMM-2020), Moscow, Russian, April 13-17, 2020

bobrikova-ev@rudn.ru (E. V. Bobrikova); aaplatonova@list.ru (A. A. Platonova); sshorgin@ipiran.ru (S. Ya. Shorgin); gaydamaka-yuv@rudn.ru (Y. V. Gaidamaka)
0000-0002-7704-5827 (E. V. Bobrikova); 0000-0003-0571-1496 (A. A. Platonova); 0000-0001-5261-0159 (S. Ya. Shorgin); 0000-0003-2655-4805 (Y. V. Gaidamaka)

Therefore, it is extremely important for network operators to use the capabilities of the new 5G networks and review business behavior. The Next Generation Mobile Networks (NGMN) Alliance assigned to Network Slicing a decisive role in the further development of 5G networks and in the impact on updating the interaction of operators with business [1, 2]. Network Slicing allows service providers to create virtual end-to-end networks, adapted to application requirements (Figure 1).

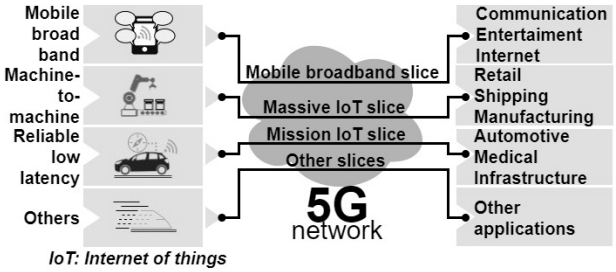

Figure 1: 5G network slicing

Using of Network Slicing allows operators to provide parts of their networks for the specific use cases of customers, for example, a smart home, Internet of things (IoT) factory, connected car, smart energy grid. Network Slicing allows to place on one physical medium with its own infrastructure various end-to-end logical networks called network slices.

Slices manage the sets of virtualized communication resources, network elements. The Network Slicing architecture can be considered as consisting of two blocks, one is for the actual implementation of the slice, and the other is for the control and configuration of a slice (Figure 2).

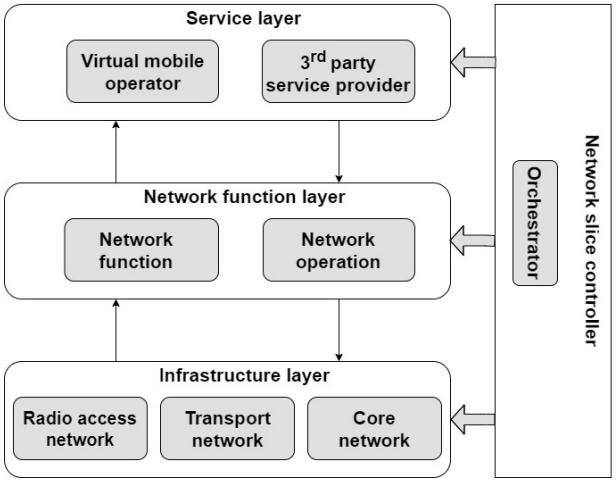

Figure 2: Network Slicing architecture

The first block is designed as a layered architecture. It consists of three levels: service layer, network function layer, infrastructure layer. Each layer contributes to the definition and deployment of the slice. The second block is implemented as a centralized network element. This is usually a controller of a network slice. The controller monitors and manages functionality between the three layers to effectively coordinate the existence and interaction of several slices.

In recent years wireless network slicing has become a central research topic to address the challenges of ever-increasing network traffic. Sharing of wireless infrastructure is widely discussed in various literature.

The paper [3] gives a rigorous mathematical formulation of the scheduling problem with multiple operators so called Generalized Resource Sharing (GRS). Authors give deep insight in the most important parameters of this scheduling and analytical and numerical characteristics of the impact of these parameters on rate-dependent utilities. Most of these results are valid for an arbitrary number of users and operators.

The paper [4] proposes the concept of Multi-Operator Scheduling (MOS). This approach allows to exchange sharing guarantees for spectral efficiency at the Base Station (BS). In addition, authors move on to a more general problem so called Anticipatory Multi-Operator Scheduling (AMOS).

It should be noted, that Network Slicing creates new problems that need to be addressed. They must be considered in order to be accepted in practice. One of the solutions is the concept of a network slice broker that acts as arbitration entity. The broker must be responsible for the meeting of the heterogeneous requirements of slices from tenants while guaranteeing the most efficient use of infrastructure resources. The paper [5] is based on the concept of brokers and develops an online network slice (ONETS) brokering solution, corresponding to the development of the new 3GPP Network Slicing architecture. The goal of ONETS solution is to develop an effective online network slice broker. It analyzes the past information about network slices and maximizes the gains of network slice resources multiplexing.

In [6] authors propose a general scheme, that describes the management of radio resource for Network Slicing; a market mechanism, that governs the allocation of radio resources for slices and an economic game, that allows to evaluate the strategies of tenant behavior at Nash equilibrium. The work [6] focuses on development a mechanism, that allows tenants to engage in a dynamic resource market. The mechanism allows tenants to optimize their solutions according to the current state of the network.

In [7] the statement of the problem considers the economic issue of the network that arises in wireless Network Slicing. The issue includes cash profit for infrastructure providers (InP) in terms of strategies for the efficient allocation of resources for several associated operators and the economic interaction of mobile virtual network operators (MVNOs) and their users. The work [7] focuses on the two-level resource allocation problem to maximize individual and total valuation of MVNOs. Here, the most important issue is the allocation of resources between MVNOs with fairness guarantee. To solve the aforementioned problems, associated with the resource allocation in wireless Network Slicing, an effective resource allocation system, using generalized Kelly mechanism (GKM), is built in [7]. The concept of GKM is based on works [8, 9]. A number of articles are known, for example [10, 11], where the problems of network slicing are solved by the methods of queuing theory [12].

Static resources distribution at each network slice does not always provide the proper quality

of service and the efficient using of the resources. The reason is in the stochastic behavior of wireless channel and the stochastic fluctuations of network traffic. To overcome these difficulties, it is proposed to use the mechanism of dynamic resource distribution [13]. With this mechanism network slices dynamically share network resources, and it becomes possible to ensure the specific requirements of slices. In this paper it is proposed a general approach, based on the introduction of an utility function, which depends on throughput and latency.

2. Main problem description

The statement of the problem and the approach to solving the problem are based on [13]. We consider a situation where a single mobile network operator (MNO) controls the downlink of a base station scheduler, whose wireless physical resources are used mutually by different network slices. Let S be the set of created network slices or tenants, since in this paper it is assumed, that the tenant controls a single slice. Let K be the set of users in the system, K_s be the subset of active users of slice $s \in S$. Each tenant $s \in S$ sets its slice requirements, which are transmitted to the base station scheduler in the form of Key Performance Indicators (KPIs), including latency and throughput.

The sharing of radio resources is modeled as a queuing system (QS), considered in discrete time. Packets arrive randomly to the base station scheduler at the beginning of each time interval (slotted time) n, $n \in N$. The arrival of packets is distributed according to Poisson's law, where λ_s is the average arrival rate of the incoming packet stream of slice s. Let b_s be the length of the packet of slice s, $D_k[n]$ be the total number of packets, arriving in the system for user k at the time slot n. Let $z_k[n]$ be the total number of served packets of user k at time slot n. It is considered that the packet is received successfully, when the total number of transmitted bits is equal to the size of the packet.

The base station scheduler allocates one buffer of infinite length for each network slice. The parameter $Q_s[n] \in \{0, 1\}$ shows the state of the s-th buffer: whether it is empty or busy at the time slot n. It is assumed, that packets are served according to the discipline First-In-First-Out (FIFO) in each buffer. The reason is that each slice defines a unique type of service, and therefore all packets are handled the same within one slice. Packages, belonged to different buffers, are served on the base of a scheduling policy, that ensures customization and differentiation of slices. To implement this statement, let $\mathbb{O}$ be the set of network performance indicators, such as, average user throughput, minimum latency. For each element $o \in \mathbb{O}$ let's define a utility function in a general form: $U_s^o(f^o(x_s), \beta_s^o)$, $\forall o \in \mathbb{O}$, $\forall s \in S$, where β_s^o is the network slice's s requirement for the specific performance indicator $o \in \mathbb{O}$, and $f^o(x_s)$ is a function, that determines the resources allocated to the network slice s for all its users x_s. Here

$$x_s = \sum_{n \in N} \sum_{k \in K_s} x_k[n]$$

and $x_k[n]$ is the share of resources, allocated by the scheduler to user k at time slot n. At last, it is assumed that the scheduler has complete information about the channel, and let $r_k[n]$ be the maximum achievable rate of user k at time slot n.

It is assumed that each network slice determines its own specific service, for which the tenant requires specially designed and customized network configuration. Network slice

customization is modeled using piece-wise linear utility functions, which display the achieved network performance indicators, based on requirements, established by tenants.

In this paper it is considered two indicators of network performance: latency and throughput, and a utility function is constructed for each of them.

Latency utility function. We define the latency for each packet as the waiting time of a packet in the buffer before this packet is transmitted. Now we define the maximum latency:

$$L_s^{\max} = \max_{k \in K_s} \{(n-i),\ \forall i \leqslant n\ :\ z_k[n] = D_k[i]\},\ \forall s \in S,$$

which describes the maximum packet latency for a slice s. Next we define the average latency:

$$L_s^{ave} = \frac{1}{|K_s|} \sum_{k \in K_s} \frac{1}{D_k} \sum_{d \in D_k} l_d,\ \forall s \in S,$$

where l_d is the latency of one packet d and D_k is the total number of packets arrived in the system to user k. L_s^{ave} determines the average latency for all users of the slice s.

For $L_s^{\max}$ and L_s^{ave} we define a utility function:

$$\hat{U}_s^L(y, \tau_s) = \begin{cases} U_{tar}^L, & \text{if } y \leqslant \tau_{tar} \\ U_{tar}^L - \dfrac{(U_{tar}^L - U_{\min}^L)(y - \tau_{tar})}{\tau_{\max} - \tau_{tar}}, & \text{if } \tau_{tar} \leqslant y \leqslant \tau_{\max} \\ U_{\min}^L, & \text{if } y \geqslant \tau_{\max}, \end{cases} \tag{1}$$

where y is the latency variable, that is either $L_s^{\max}$ or L_s^{ave}; τ_s is the latency requirement for the network slice s.

It is assumed, that each network slice defines its interval of the required latency, that is, the target latency τ_{tar} and the maximum allowable latency $\tau_{\max}$, therefore $\tau_s = \{\tau_{tar}, \tau_{\max}\}$. $U_{\min}^L$ is the minimum value of the utility function, U_{tar}^L is the maximum value of the utility function. The general view of the latency utility function is in Figure 3. The values of the function $\hat{U}_s^L$ are calculated according to the Algorithm 1.

Data: y
Result: $\hat{U}_s^L(y, \tau_s)$
if $y \leqslant \tau_{tar}$ **then**
  $\hat{U}_s^L(y, \tau_s) = U_{tar}^L$
else if $\tau_{tar} < y \leqslant \tau_{\max}$ **then**
  $\hat{U}_s^L(y, \tau_s) = U_{tar}^L - \frac{(U_{tar}^L - U_{\min}^L)(y - \tau_{tar})}{\tau_{\max} - \tau_{tar}}$
else
  $\hat{U}_s^L(y, \tau_s) = U_{\min}^L$
end

Algorithm 1: The values of $\hat{U}_s^L$.

Based on the definitions given above, the overall latency utility function is defined as follows:

$$U_s^L = \delta_s \cdot \hat{U}_s^L(L_s^{\max}, \tau_s) + (1 - \delta_s) \cdot \hat{U}_s^L(L_s^{ave}, \tau_s),\ \forall s \in S,$$

where δ_s is the weight coefficient specified by the tenant. The coefficient δ_s determines the priority of the network slice using the latency utility function in terms of $L_s^{\max}$ or L_s^{ave}.

Throughput utility function. We define the total user throughput for one slice:

$$R_s = \frac{1}{T_s^{ACT}} \sum_{k \in K_s} z_k[N] \cdot b_k, \ \forall s \in S,$$

where $z_k[N]$ is the total number of packets, transferred to the user k, and T_s^{ACT} is the total time, during which buffer is active for packets transmission. The throughput utility function is defined as follows:

$$U_s^T(y, \rho_s) = \begin{cases} U_{tar}^T, \text{ if } y \geqslant \rho_{tar}, \\ U_{tar}^T - \dfrac{(U_{tar}^T - U_{\min}^T)(y - \rho_{tar})}{\rho_{\min} - \rho_{tar}}, \text{ if } \rho_{tar} \geqslant y \geqslant \rho_{\min}, \\ U_{\min}^T \dfrac{y - \rho_{zero}}{\rho_{\min} - \rho_{zero}}, \text{ if } \rho_{\min} \geqslant y \geqslant \rho_{zero}, \\ 0, \text{ if } y \leqslant \rho_{zero}, \end{cases} \tag{2}$$

where $y = R_s$ is the aggregate throughput of the user of slice s and $\rho_s = \{\rho_{tar}, \rho_{\min}, \rho_{zero}\}$ are the throughput requirements; ρ_{zero} — the basic bit-rate for each slice; $\rho_{\min}$ — the minimum guaranteed bit-rate, which is necessary to provide the standard quality of service; $U_{\min}^T$ — the value of the utility function corresponding to $\rho_{\min}$; ρ_{tar} — the bit-rate, which is necessary to provide a high quality of service; U_{tar}^T is the maximum value of the utility function corresponding to ρ_{tar}. The general view of the throughput utility function is in Figure 4. The values of the function $\hat{U}_s^L$ are calculated according to the Algorithm 2.

Data: y
Result: $U_s^T(y, \rho_s)$
if $y \leqslant \rho_{zero}$ **then**
  $U_s^T(y, \rho_s) = 0$
else if $\rho_{zero} < y \leqslant \rho_{\min}$ **then**
  $U_s^T(y, \rho_s) = U_{\min}^T \frac{y - \rho_{zero}}{\rho_{\min} - \rho_{zero}}$
else if $\rho_{\min} < y \leqslant \rho_{tar}$ **then**
  $U_s^T(y, \rho_s) = U_{tar}^T - \frac{(U_{tar}^T - U_{\min}^T)(y - \rho_{tar})}{\rho_{\min} - \rho_{tar}}$
else
  $U_s^T(y, \rho_s) = U_{tar}^T$
end

Algorithm 2: The values of U_s^T.

The resource allocation problem. The scheduler assigns physical resources to users based on the requirements of the network slice, using the latency utility function U_s^L and the throughput utility function U_s^T, defined above. In addition, we introduce a specific parameter $\underline{\alpha}_s$ for a network slice in order to start the mechanism that allows tenants themselves to determine the weights of the corresponding utility function.

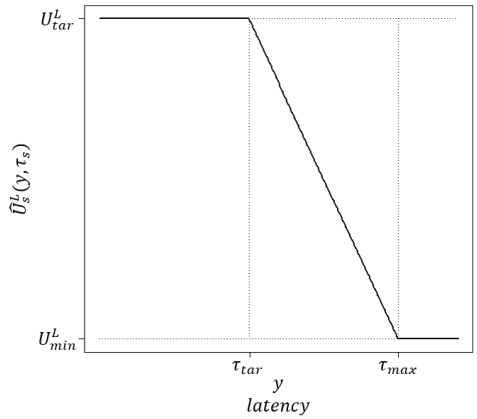

Figure 3: The general view of the latency utility function

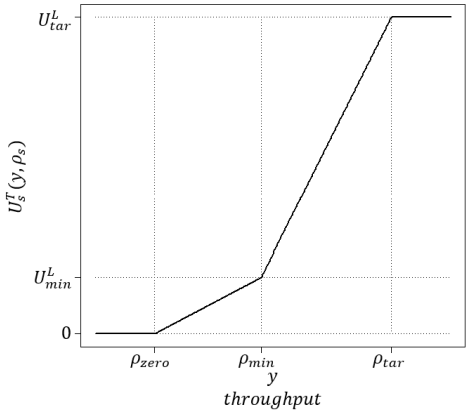

Figure 4: The general view of the throughput utility function

Let us formulate the resource allocation problem, oriented to network slices, as an optimization problem [13, 14] in the following statement

$$\max \sum_{s \in S} \underline{U}_s \cdot \underline{\alpha}_s,$$

where

$$\sum_{k \in K} x_k[n] \leqslant 1, \ \forall n \in N, \quad z_k[n] \leqslant D_k[n-1], \ \forall k \in K, \forall n \in N,$$

$$\sum_{i=1}^{n} x_k[i] \cdot r_k[i] \leqslant D_k[n] \cdot b_k, \ \forall k \in K, \forall n \in N, \quad \sum_{i=1}^{n} x_k[i] \cdot r_k[i] \geqslant z_k[n] \cdot b_k, \ \forall k \in K, \forall n \in N,$$

$$Q_s[n] = \begin{cases} 0, \text{ if } D_k[n] = z_k[n] \\ 1, \text{ otherwise} \end{cases}, \quad \forall n \in N, \forall k \in K_s, \forall s \in S.$$

Here the optimal solution provides maximum weighted sum of the utility functions of the slices for all slices of the network. The problem is formulated for both performance indicators, that is, $\underline{U}_s = \{U^L_s, U^T_s\}$, $\underline{\alpha}_s = \{\alpha^L_s, \alpha^T_s\}$. The first constraint ensures, that the scheduler does not assign more resources than those are available in network at every time slot n. The second constraint shows, that the packet can be considered successfully transmitted only at the end of the time slot or at the beginning of the next time slot. The third constraint ensures, that the number of bits transmitted cannot be more than the total number of bits received in the system at one time slot. The fourth constraint updates the total number of received packets at each time slot n, taking into account the total number of received bits. Finally, the fifth constraint determines the state of the buffer.

The proposed formulation of the resource allocation problem always guarantees the maximization of the utility function for each network slice. Therefore, in the case of a sufficient amount of resources in the system, we can assume that each network slice reaches maximum of its utility function. However, the mobile network operator (MNO) must be ready for situations when resources are not enough, for example, due to network congestion. Therefore it is assumed

that the tenant can monitor the performance of the slice in real time and can change its priority indicators in order to scale the utility function and get various network performance indicators.

By introducing the specific parameter $\underline{\alpha}_s$ for the slice, we enable a tenant to configure and differentiate the network slice. This becomes especially impotant in case of congestion of the network, since MNO has to make decisions how to deal with slices, when it is known beforehand that the execution of all the requirements may not be possible. In this sense, the adjusting of $\underline{\alpha}_s$ allows tenants:

— to prioritize a set of network performance indicators within a single slice (*slice customization*). That is, whenever MNO is not be able to fulfill all the requirements for the slice, resources is allocated to maximize utility of the indicator with higher priority.

— to prioritize slices (*slice differentiation*). In this case, when the MNO is not able to provide maximum value of the utility function for all slices, parameter $\underline{\alpha}_s$ indicates the most critical slices, that require higher priority.

Note, that the utility function is defined for both latency and throughput. Therefore, it is possible to consider services with different parameters of latency and throughput, using the capabilities of Network Slicing. Let's consider the following services: TI (Tactile Internet), eMBB (enhanced Mobile BroadBand), mMTC (massive Machine Type Communication), cMTC (critical Machine Type Communication) [13]. These services are supported by 5G, the principles of Network Slicing are applicable to which, and for which it is possible to construct utility functions, that take into account both latency and throughput. It is shown in Figure 5 the latency utility functions for four types of the slices, namely: $\hat{U}^L_{TI}(y, \tau_{TI})$, $\hat{U}^L_{eMBB}(y, \tau_{eMBB})$, $\hat{U}^L_{mMTC}(y, \tau_{mMTC})$, $\hat{U}^L_{cMTC}(y, \tau_{cMTC})$, based on (1).

TI and cMTC services relate to URLLC (Ultra-Reliable Low Latency Communication) applications. These two slices are the most critical applications in terms of latency. Latency requirements can even have values below 1 ms [13]. These applications also require high throughput. It is assumed that eMBB and mMTC slices are more flexible with respect to latency requirements and their utility is less affected by delays in scheduling decisions. It is shown in Figure 6 the throughput utility functions for four types of the slices, namely: $U^T_{TI}(y, \tau_{TI})$, $U^T_{eMBB}(y, \tau_{eMBB})$, $U^T_{mMTC}(y, \tau_{mMTC})$, $U^T_{cMTC}(y, \tau_{cMTC})$, based on (2).

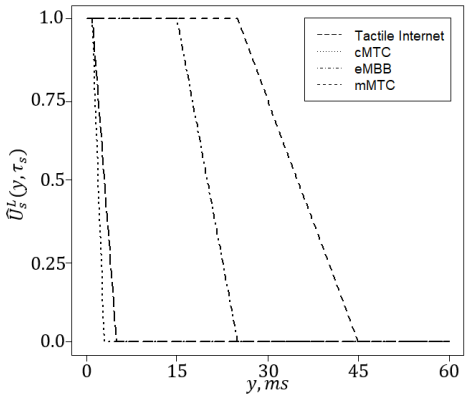

Figure 5: Latency utility functions for different services: TI, eMBB, mMTC, cMTC

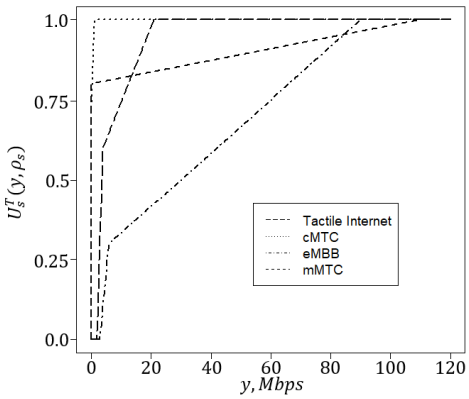

Figure 6: Throughput utility functions for different services: TI, eMBB, mMTC, cMTC

It is assumed, that for applications with weak throughput requirements, achieving a minimum guaranteed bit-rate is sufficient to provide satisfactory service, which means a high value for utility. It takes place for TI, cMTC, mMTC. On the contrary, for the eMBB slice, the utility value is set much lower, which means low provide quality. As we can see, as a result of strict latency requirements, for applications TI and cMTC, an increase in throughput leads to a decrease in latency. On the contrary, applications, which are not latency-critical, are more demanding in terms of aggregate throughput. Namely, for the mMTC slice it is assumed that the minimum guaranteed bit-rate should be provided, but the aggregate throughput can be high, given the huge number of connected devices. It is assumed for the eMBB slice, a high throughput requirement for each user, but with relatively few users at the same time active per cell.

The parameters for the utility functions in Figures 5, 6 are derived according to [13]. But we consider slightly different combinations of parameters for the presented slices.

The approach considered in this paper, using the construction of the utility function of the slice, can be recommended to MNO. For their set of services, operators will be able to evaluate the profitability of these services and determine the tariffs for users. So, in [13] a numerical analysis was performed for TI, cMTC, eMBB, mMTC slices. The scheduling tactics depend mainly on two factors: how utility functions are defined for the different types of slices and how tenants determine weights by choosing the parameter $\underline{\alpha}_s$. Some recommendations are given. In order to maximize the utility of the latency-critical slices, it is necessary to schedule user service, as soon as the packet arrives at the buffer, regardless of the state of their channel. This method gives the highest priority to such users and does not allow the scheduler to make more effective scheduling decisions. On the contrary, for slices that are oriented to high throughput, the state of the user channel may also influence the increase in utility.

3. Conclusions

In this paper we propose an algorithm for the dynamic sharing of network resources by network slices. The utility function of the network slice is described, which allows to customize the behavior of various types of slices. Differentiation between tenants is achieved through change in specific parameters of the slices. These parameters, in turn, dynamically change the view of the utility function of the slice. In the future it is planned to study in detail the effect of parameter changes in some slices of the network on service latency in other slices. It is planned to investigate the minimum value of the utility function, corresponding to the minimum guaranteed bit-rate for each slice. It is planned to consider the impact of changing of the parameters of the task in connection with the interaction between tenants and mobile network operators.

Acknowledgments

The work is supported by RUDN Program «5-100» and by RFFR in the framework of the scientific projects № 18-07-00576, 19-07-00933. The authors thank Natalia Yarkina for the materials that were used in the Introduction.

References

[1] 3GPP, Study on management and orchestration of network slicing for next generation network, 2017. URL: https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=3091.

[2] D. Sattar, A. Matrawy, Optimal slice allocation in 5g core networks, IEEE Networking Letters 1 (2019) 48–51. doi:10.1109/LNET.2019.2908351.

[3] I. Malanchini, S. Valentin, O. Aydin, An analysis of generalized resource sharing for multiple operators in cellular networks, in: 2014 IEEE 25th Annual International Symposium on Personal, Indoor, and Mobile Radio Communication (PIMRC), IEEE, 2014, pp. 1157–1162. doi:10.1109/PIMRC.2014.7136342.

[4] I. Malanchini, S. Valentin, O. Aydin, Wireless resource sharing for multiple operators: Generalization, fairness, and the value of prediction, Computer Networks 100 (2016) 110–123. doi:10.1016/j.comnet.2016.02.014.

[5] V. Sciancalepore, L. Zanzi, X. Costa-Perez, A. Capone, Onets: online network slice broker from theory to practice, arXiv preprint arXiv:1801.03484 (2018).

[6] A. Lieto, E. Moro, I. Malanchini, S. Mandelli, A. Capone, Strategies for network slicing negotiation in a dynamic resource market, in: 2019 IEEE 20th International Symposium on "A World of Wireless, Mobile and Multimedia Networks" (WoWMoM), IEEE, 2019, pp. 1–9. doi:10.1109/WoWMoM.2019.8792999.

[7] Y. K. Tun, N. H. Tran, D. T. Ngo, S. R. Pandey, Z. Han, C. S. Hong, Wireless network slicing: Generalized kelly mechanism-based resource allocation, IEEE Journal on Selected Areas in Communications 37 (2019) 1794–1807. doi:10.1109/JSAC.2019.2927100.

[8] F. P. Kelly, A. K. Maulloo, D. K. Tan, Rate control for communication networks: shadow prices, proportional fairness and stability, Journal of the Operational Research society 49 (1998) 237–252. doi:10.1057/palgrave.jors.2600523.

[9] F. Kelly, Charging and rate control for elastic traffic, European transactions on Telecommunications 8 (1997) 33–37. doi:10.1002/ett.4460080106.

[10] M. Vincenzi, E. Lopez-Aguilera, E. Garcia-Villegas, Maximizing infrastructure providers' revenue through network slicing in 5g, IEEE Access 7 (2019) 128283–128297. doi:10.1109/ACCESS.2019.2939935.

[11] I. Vilà, O. Sallent, A. Umbert, J. Pérez-Romero, An analytical model for multi-tenant radio access networks supporting guaranteed bit rate services, IEEE access 7 (2019) 57651–57662. doi:10.1109/ACCESS.2019.2913323.

[12] Y. Gaidamaka, A. Pechinkin, R. Razumchik, K. Samouylov, E. Sopin, Analysis of an mg 1 r queue with batch arrivals and two hysteretic overload control policies, International Journal of Applied Mathematics and Computer Science 24 (2014) 519–534. doi:10.2478/amcs-2014-0038.

[13] A. Lieto, I. Malanchini, A. Capone, Enabling dynamic resource sharing for slice customization in 5g networks, in: 2018 IEEE Global Communications Conference (GLOBECOM), IEEE, 2018, pp. 1–7. doi:10.1109/GLOCOM.2018.8647249.

[14] H. A. Taha, Operations research: an introduction, seventh edition, Williams Publishing House, 2005.