



Scenario of priority service for elastic traffic within the network slicing framework

Anastasia A. Tazaeva¹, Ekaterina V. Markova^{1,2}

¹ *RUDN University, 6 Miklukho-Maklaya St, Moscow, 117198, Russian Federation*

² *Federal Research Center “Computer Science and Control” of the Russian Academy of Sciences, 44 Vavilova St, bldg 2, Moscow, 119333, Russian Federation*

(received: May 29, 2025; revised: February 20, 2026; accepted: March 3, 2026)

Abstract. *Background.* Network slicing (NS) enables multiple logical slices over shared infrastructure but faces challenges in managing heterogeneous traffic (streaming vs. elastic) with conflicting QoS requirements. *Purpose.* This paper proposes a radio admission control (RAC) scheme with priority service interruption for a two-slice 5G network to balance GBR for streaming traffic and efficient non-GBR elastic traffic service. *Method.* A two-dimensional Markov model is developed. Streaming traffic requires fixed resources; elastic traffic has state-dependent service rate. A full interruption scenario is implemented where elastic traffic preempts streaming traffic within the second slice's guaranteed capacity. Stationary probabilities are obtained numerically. *Results.* As streaming traffic arrival rate increases, elastic traffic interruption probability rises sharply then stabilises. Mean numbers of served requests converge to guaranteed capacities. Elastic traffic blocking remains nearly constant. *Conclusions.* The proposed RAC scheme captures the streaming-elastic trade-off in sliced networks. The full interruption mechanism prioritises elastic traffic under high load. The model provides quantitative guidelines for tuning guaranteed capacities in 5G/6G systems.

Key words and phrases: 5G, network slicing, mobile networks, QoS, GBR services, non-GBR services, key performance indicators, priority service, service interruption, radio resource management, resource allocation

For citation: A. A. Tazaeva, E. V. Markova “Scenario of priority service for elastic traffic within the network slicing framework,” *Discrete and Continuous Models and Applied Computational Science*, vol. 34, no. 3, pp. 334–344, 2026. DOI: 10.22363/2658-4670-2026-34-3-334-344. EDN: RRPVBC.

1. Introduction

The rapid advancement of fifth-generation (5G) networks has significantly increased the demand for enhanced Quality of Service (QoS) from service providers. Organizations such as 3GPP and the GSM Association emphasize that 5G networks must support multiaccess capabilities, maintain low and predictable latency with minimal packet loss, and ensure high network resilience [1–3]. These stringent requirements have led to the development of flexible and adaptive network architectures capable of dynamically addressing these challenges [4–7].

One of the fundamental technologies that allows for this flexibility is Network Slicing (NS) [8–11]. NS allows mobile network operators to create multiple logically isolated and customizable network slices within a shared physical infrastructure, such as a base station (BS). These slices, referred to as Network Slice Instances (NSIs), can be dynamically deployed and removed based on real-time technical requirements, facilitating autonomous resource allocation without direct user involvement. Furthermore, NS enables service providers to manage QoS efficiently, enhancing fault tolerance and overall network reliability [12–15].

Each NSI is typically designed to support a specific category of service, such as best effort with minimum guarantee (BG), guaranteed bit rate (GB), or best effort (BE). BG and BE services generate elastic traffic with variable service duration, while GB services generate streaming traffic with fixed duration [16, 17].



This paper proposes a model for the joint servicing of two types of traffic, streaming and elastic with minimum guarantee, within a 5G network utilizing the NS framework. In addition, it evaluates key performance metrics to assess the efficiency of the model.

The remainder of this paper is structured as follows. Section 2 introduces the radio admission control (RAC) scheme for a 5G network, formulated as a queueing system (QS) managing both streaming and elastic traffic. The analysis includes transition intensity diagrams, an infinitesimal generator, and calculations of key performance measures. Finally, Section 3 presents the conclusions.

2. System model

Consider the operation of a base station (BS) cell within a network with a total capacity of C [capacity units, c.u.]. This capacity is distributed between two NSIs, each with capacities C_1 and C_2 , respectively, where $\sum_{k=1}^2 C_k \geq C$. The capacity C_k allocated to NSI k consists of two parts: a guaranteed capacity Q_k and a shared capacity available to other NSIs, expressed as $C_k - Q_k$, for $k = 1, 2$. The guaranteed capacities satisfy the condition $\sum_{k=1}^2 Q_k \leq C$, where $0 < Q_k \leq C_k \leq C$, for $k = 1, 2$.

We assume that the first NSI is for services generating streaming traffic (first-type requests), while the second NSI accommodates services generating elastic traffic (second-type requests). The arrival process for service requests follows a Poisson distribution with an arrival rate λ_k , where $k = 1, 2$. The service time for first-type requests is exponentially distributed with a mean of μ_1^{-1} . In contrast, the duration of the service for second-type requests depends on the system load and the average file length Θ .

Each first-type request requires b_1 c.u., constrained by $b_1 \leq Q_1$. Similarly, each second-type request requires a minimum allocation of $b_2^{\min}$ c.u., where $b_2^{\min} \leq Q_2$. For simplicity, we set $b_1 = b_2^{\min} = 1$.

Let n_k denote the number of active requests in NSI k ($k = 1, 2$) when the system is in state $n = (n_1, n_2)$. Given the elastic nature of second-type traffic, the number of occupied capacity units is defined as

$$b_2(n_1, n_2) = \frac{\min(C - b_1 n_1, C_2)}{n_2},$$

where $b_2^{\min} \leq b_2(n_1, n_2) \leq C_2$. The corresponding service time for second-type requests is determined by

$$\mu_2^{-1}(n_1, n_2) = \frac{n_2 \Theta}{b_2(n_1, n_2)}.$$

The key parameters are summarized in Table 1.

Now, let us analyze the request admission process in the system. When a new first-type request arrives, the following conditions apply:

- The request is admitted if the number of ongoing first-type requests is below the maximum allowable limit, given by $N_1^{\max} = C_1/b_1$, and there are sufficient available resources, specifically at least b_1 capacity units.
- If the number of first-type requests is below the threshold supported by the guaranteed capacity, i.e., $N_1^g = Q_1/b_1$, but available resources are insufficient (less than b_1), the request can still be accepted by interrupting a second-type request, provided that at least one second-type request exists within the guaranteed capacity Q_1 .
- If neither of the above conditions is met, the request is blocked.

The radio admission control mechanism for second-type requests is structured as follows:

- The request is admitted if the number of ongoing second-type requests is below the maximum permissible limit, given by $N_2^{\max} = C_2/b_2^{\min}$, and there are sufficient available resources, specifically at least $b_2^{\min}$ capacity units.
- If the number of second-type requests is below the threshold supported by the guaranteed capacity, i.e., $N_2^g = Q_2/b_2^{\min}$, but available resources are insufficient (less than $b_2^{\min}$), the request can still be accepted by interrupting a first-type request, provided that at least one first-type request exists within the guaranteed capacity Q_2 .

Table 1

Summary of key parameters

Notation	Description
C	The BS's total network capacity (capacity units, c.u.).
C_k	The k -th NSI's overall network capacity, $k \in \{1, 2\}$, $C_1 + C_2 \geq C$ (c.u.).
Q_k	The k -th NSI's guaranteed network capacity, $Q_k \leq C_k$, $Q_1 + Q_2 \leq C$ (c.u.).
λ_k	The arrival rate of requests in k -th NSI, $k \in \{1, 2\}$ (requests per unit of time).
μ_1^{-1}	The average service time of one request at the first NSI (t.u.).
Θ	The average length of the transmitted file (c.u. $\times$ t.u.).
b_1	The number of resource units required to serve each request of the first type (c.u.).
$b_2^{\min}$	The minimal number of resource units required to serve each request of the second type (c.u.).
n_k	The number of requests in k -th NSI, $k \in \{1, 2\}$ (—).
$N_k^{\max}$	The maximum number of requests in k -th NSI, $N_1^{\max} = C_1/b_1$, $N_2^{\max} = C_2/b_2^{\min}$ (—).
N_k^g	The maximum number of requests that can be served using guaranteed network capacity of k -th NSI (—).
$b_2(n_1, n_2)$	The number of resource units available for servicing each request of the second type in the state (n_1, n_2) , $b_2^{\min} \leq b_2(n_1, n_2) \leq C_2$ (c.u.).

- If neither of these conditions is met, the second-type request is blocked.

In cases where resources are insufficient, service interruption for first-type requests can occur through two mechanisms:

1. If the number of ongoing second-type requests is below the guaranteed threshold, i.e., $n_2 < N_2^g$, a first-type request is interrupted to free up the minimum required resources, $b_2^{\min}$, necessary for accommodating a second-type request.
2. If the number of ongoing second-type requests remains below the guaranteed threshold, i.e., $n_2 < N_2^g$, all second-type requests utilizing the guaranteed capacity Q_2 are interrupted.

This study focuses on the second scenario. The system operation is illustrated schematically in Figure 1.

3. Mathematical model

The system behavior is described by a two-dimensional Markov process (MP) $\{(X_1(t), X_2(t)), t > 0\}$, where $X_k(t)$ represents the number of k type requests in the system at time t , $k = 1, 2$, over the state space $\mathbf{X}$:

$$\mathbf{X} = \{(n_1, n_2) : 0 \leq n_1 \leq N_1^{\max} \wedge 0 \leq n_2 \leq N_2^{\max} \wedge n_1 b_1 + n_2 b_2^{\min} \leq C \wedge n_2 b_2(n_1, n_2) \cdot \mathbf{I}(n_2 \geq 1) \geq Q_2\},$$

where I is a characteristic function

$$\mathbf{I}(P) = \begin{cases} 1, & \text{if the condition is satisfied,} \\ 0, & \text{if the condition is not satisfied.} \end{cases}$$

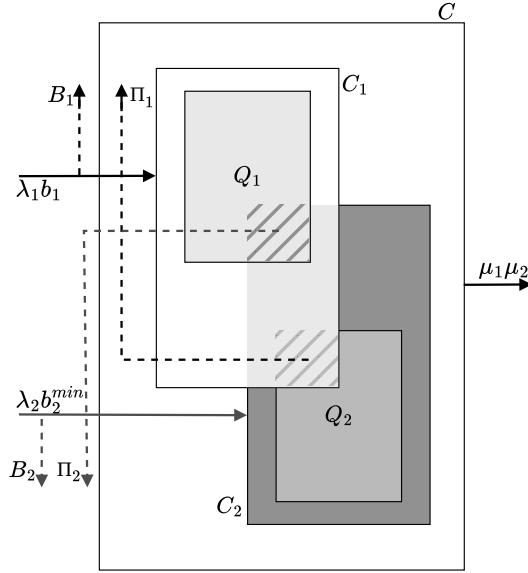


Figure 1. System model representation

The state space $\mathbf{X}$ includes few subsets: interruption subset $\mathbf{V}_i^k$, admission subset $\mathbf{S}_k$, and blocking subset $\mathbf{B}_k$, where $i, k \in \{1, 2\}$, $i \neq k$. Based on the sets, the transition intensity diagram can be represented in the following form (Figure 2).

The interruption subset $\mathbf{V}_1^2$, $\mathbf{V}_2^1$, represents the system states where the service of arriving k -type request has priority over the service of i -type request (in other words, states where the service of i -type request can be interrupted to service one arriving k type request):

$$\mathbf{V}_1^2 = \{(n_1, n_2) \in \mathbf{X} : n_2 = 0 \wedge n_2 b_2(n_1, n_2) < Q_2\},$$

$$\mathbf{V}_2^1 = \{(n_1, n_2) \in \mathbf{X} : n_1 < N_1^g \wedge n_1 b_1 + n_2 b_2^{\min} = C\}.$$

The admission subset $\mathbf{S}_k$, $k = 1, 2$, represents the system states where k -type request will be accepted for service in the k -th NSI:

$$\begin{aligned} \mathbf{S}_1 = \mathbf{V}_2^1 \cup \{ & (n_1, n_2) \in \mathbf{X} : (0 \leq n_1 < N_1^{\max} \wedge (n_1 + 1)b_1 + \\ & + n_2 b_2^{\min} \leq C \wedge n_2 = 0) \vee (0 \leq n_1 < N_1^{\max} \wedge (n_1 + 1)b_1 + \\ & + n_2 b_2^{\min} \leq C \wedge n_2 > 0 \wedge n_2 b_2(n_1, n_2) > Q_2) \}, \end{aligned}$$

$$\mathbf{S}_2 = \mathbf{V}_1^2 \cup \{(n_1, n_2) \in \mathbf{X} : 0 \leq n_2 < N_2^{\max} \wedge n_1 b_1 + (n_2 + 1)b_2^{\min} \leq C\}.$$

The blocking subset $\mathbf{B}_k$, $k = 1, 2$, represents the system states where arriving k -type requests will be blocked by the system due to insufficient available resources in the k -th NSI:

$$\begin{aligned} \mathbf{B}_1 = \{ & (n_1, n_2) \in \mathbf{X} : n_1 = N_1^{\max} \vee (N_1^g \leq n_1 < N_1^{\max} \wedge \\ & \wedge (n_1 + 1)b_1 + n_2 b_2^{\min} > C) \vee (n_2 > 0 \wedge n_2 b_2(n_1 + 1, n_2) < Q_2) \}. \end{aligned}$$

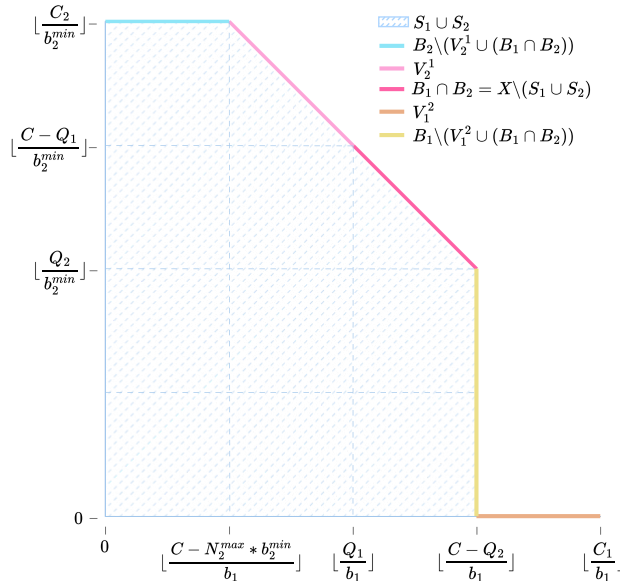


Figure 2. Scheme of the transition intensity diagram

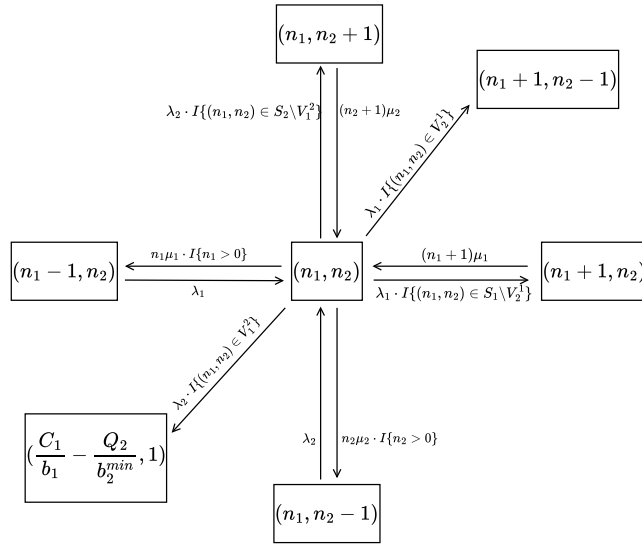
$$\mathbf{B}_2 = \left\{ (n_1, n_2) \in \mathbf{X} : (n_2 = N_2^{\max}) \vee \right. \\ \left. \vee (N_2^{\xi} \leq n_2 < N_2^{\max} \wedge n_1 b_1 + (n_2 + 1) b_2^{\min} > C) \right\}.$$

According to the diagram of transition intensities (Figure 3), the considered process is described by the following system of equilibrium equations:

$$\begin{aligned} & p(n_1, n_2) \cdot (\lambda_2 \cdot \mathbf{I}\{(n_1, n_2) \in \mathbf{S}_2 / \mathbf{V}_1^2\} + \lambda_1 \cdot \mathbf{I}\{(n_1, n_2) \in \mathbf{V}_2^1\} \\ & + \lambda_1 \cdot \mathbf{I}\{(n_1, n_2) \in \mathbf{S}_1 / \mathbf{V}_2^1\} + \lambda_2 \cdot \mathbf{I}\{(n_1, n_2) \in \mathbf{V}_1^2\} \\ & + n_2 \mu_2 \cdot \mathbf{I}\{n_2 > 0\} + n_1 \mu_1 \cdot \mathbf{I}\{n_1 > 0\}) \\ & = [\lambda_1 \cdot p(n_1 - 1, n_2) + \lambda_2 \cdot p(n_1, n_2 - 1) + (n_1 + 1) \mu_1 \cdot p(n_1 + 1, n_2) \\ & + (n_2 + 1) \mu_2 \cdot p(n_1, n_2 + 1)], \end{aligned}$$

where $p(n_1, n_2)$, $n \in \mathbf{X}$ — stationary probability distribution of system states.

Due to the implementation of the mechanism for interrupting the service of requests, the process $\{(X_1(t), X_2(t)), t > 0\}$, describing the considered system, is not a reversible Markov process. Therefore, the system's stationary probability distribution $\mathbf{p} = \{p(n_1, n_2), (n_1, n_2) \in \mathbf{X}\}$ can be computed using the numerical solution of the equilibrium equations system $\mathbf{p}^T \cdot \mathbf{A} = \mathbf{0}^T$, $\mathbf{p}^T \cdot \mathbf{1} = 1$, where $\mathbf{A}$ is the infinitesimal generator, and the elements $\mathbf{a}((n_1, n_2), (n'_1, n'_2))$ of this generator are determined by the following formulas:

Figure 3. The diagram of transition intensities for an arbitrary system state (n_1, n_2)

$$a((n_1, n_2), (n'_1, n'_2)) = \begin{cases} \lambda_1, n'_1 = n_1 + 1, n'_2 = n_2, (n_1, n_2) \in \mathbf{S}_1 / \mathbf{V}_2^1, \\ \lambda_1, n'_1 = n_1 + 1, n'_2 = n_2 - 1, (n_1, n_2) \in \mathbf{V}_2^1, \\ \lambda_2, n'_1 = n_1, n'_2 = n_2 + 1, (n_1, n_2) \in \mathbf{S}_2 / \mathbf{V}_1^2, \\ \lambda_2, n'_1 = \frac{C_1}{b_1} - \frac{Q_2}{b_2^{\min}}, n'_2 = 1, (n_1, n_2) \in \mathbf{V}_1^2, \\ n_1\mu_1, n'_1 = n_1 - 1, n'_2 = n_2, n_1 > 0, \\ n_2\mu_2, n'_1 = n_1, n'_2 = n_2 - 1, n_2 > 0, \\ \gamma, n'_1 = n_1, n'_2 = n_2, \\ 0, \text{otherwise,} \end{cases}$$

where

$$\gamma = -\left(\lambda_1 \cdot \mathbf{I}\{(n_1, n_2) \in \mathbf{S}_1\} + \lambda_2 \cdot \mathbf{I}\{(n_1, n_2) \in \mathbf{S}_2\} + n_1\mu_1 \cdot \mathbf{I}\{(n_1, n_2) \in \mathbf{X} : n_1 > 0\} + n_2\mu_2 \cdot \mathbf{I}\{(n_1, n_2) \in \mathbf{X} : n_2 > 0\}\right).$$

4. The main performance measures

The main performance measures of the model are the following:

- The mean number N_k of k -type requests served in the k -th NSI

$$N_k = \sum_{(n_1, n_2) \in \mathbf{X}} n_k \cdot p(n_1, n_2), \quad k = 1, 2.$$

- The service interruption probability Π_i^k of i type requests served in the i -th NSI when new k -type request arrives at the system

$$\Pi_i^k = \sum_{(n_1, n_2) \in \mathbf{V}_i^k} p(n_1, n_2), \quad i, k = 1, 2, \quad i \neq k.$$

- The accepting probability S_k of k -th type request for service at k -th NSI

$$S_k = \sum_{(n_1, n_2) \in \mathbf{S}_k} p(n_1, n_2), \quad k = 1, 2.$$

- The blocking probability B_k of k -th type request

$$B_k = 1 - S_k = \sum_{(n_1, n_2) \in \mathbf{B}_k} p(n_1, n_2), \quad k = 1, 2.$$

5. Results

Let us now proceed to the numerical analysis. As a case study of a system supporting the joint handling of streaming and elastic traffic, we could consider a 5G network that provides both enhanced Mobile Broadband (eMBB) and Ultra-Reliable Low-Latency Communication (URLLC) services. In this context, eMBB corresponds to elastic traffic, while URLLC represents streaming traffic. We examine how key performance metrics—such as the mean number of active k -type requests, service interruption probability, accepting probability, and blocking probability—vary with changes in the arrival rate, for example, first-type requests.

The parameters used in this numerical example are summarized in Table 2.

Main parameters of the numerical example

Table 2

Notation	Value
C	25
C_1	20
C_2	15
Q_1	13
Q_2	10
λ_1	$0 \rightarrow 120$
λ_2	100
μ_1	1
Θ	1.5
$b_1 = b_2^{\min}$	1

Figure 4 presents the dependence of the service interruption probability Π_i^k on the arrival rate λ_1 . The plot demonstrates that, for the first-type of requests, the interruption probability increases slightly as λ_1 grows. On the contrary, for the second-type requests the probability of interruption increases sharply at first, and then also decreases sharply with an increase in the value of λ_1 and eventually approaches a stable value. This behavior occurs due to the fact that the intensity of the second-type requests is high, and they immediately occupy almost the entire system capacity, and therefore, when first-type requests arrival at the system, the guaranteed network capacity is released and service is interrupted. However, since the number of the second-type requests that can be interrupted is limited by the threshold value of Q_1 , then there is a decrease in the interruption probability.

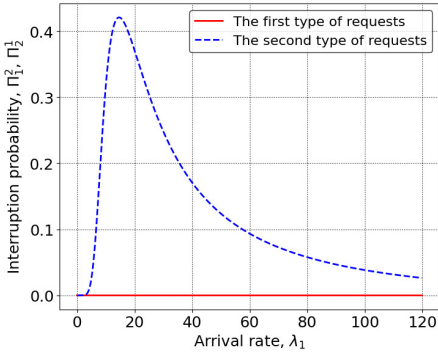


Figure 4. Dependence of the interruption probability on the arrival rate of the first-type requests

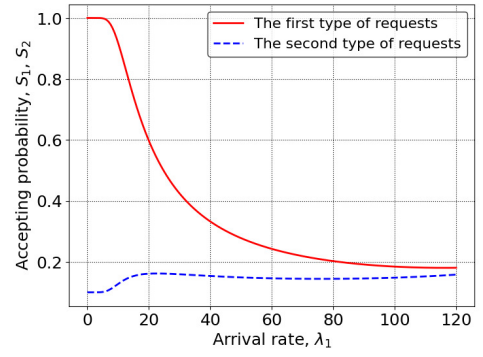


Figure 5. Dependence of the accepting probability on the arrival rate of the first-type requests

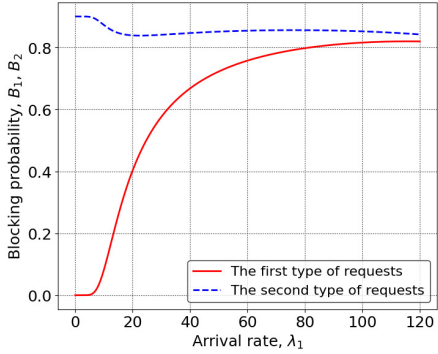


Figure 6. Dependence of the blocking probability on the arrival rate of the first-type requests

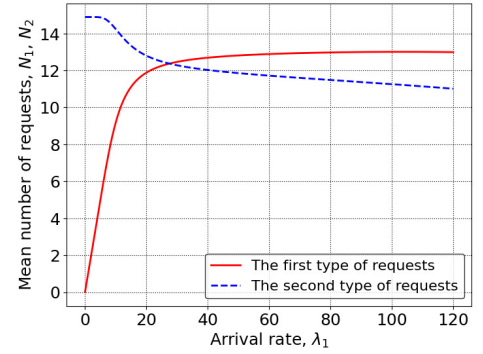


Figure 7. Dependence of the mean number of serviced requests on the arrival rate of the first-type requests

Let's consider the behavior of the accepting probability S_k for each type of requests, depending on the first-type requests arrival rate, λ_1 . As shown in Figure 5, at the initial moment of time, with a small λ_1 , the accepting probability for the first-type requests is maximum and tends to unity, since requests of each type occupy the capacity of the corresponding NSI and do not affect the QoS for each other. As the arrival rate of requests increases, the accepting probability for both types of requests decreases rapidly. However, since the arrival rate of the second-type requests, (λ_2), remains unchanged, the decrease in S_2 is less pronounced in the graph.

A similar trend is observed in Figure 6, which depicts the blocking probabilities. As λ_1 increases, the blocking probability B_1 grows correspondingly, while the blocking probability B_2 remains nearly constant due to the fixed arrival rate λ_2 .

We now turn to the most insightful graph, which illustrates the dependence of the mean number of serviced requests N_k , for $k = 1, 2$, on the arrival rate λ_1 of the first-type requests. As shown in Figure 7, when no first-type requests are present in the system, second-type requests utilize the entire portion of capacity $C_2 = 15$ allocated to them and the number of second-type requests served in the system tends to the maximum value $N_2^{\max} = 15$. As first-type requests begin to arrive, the number of serviced second-type requests N_2 gradually decreases. Under conditions of full system load, the number of serviced requests for both types converges to their respective guaranteed levels $N_1^g = 13$ and $N_2^g = 10$.

6. Discussion

The obtained results highlight the interaction between streaming and elastic traffic in a sliced 5G network. The interruption probability of elastic flows increases sharply in the low-to-medium λ_1 regime, but then stabilizes as the guaranteed resources Q_1 and Q_2 become dominant. This confirms the importance of guaranteed quotas for maintaining predictable service.

Compared to earlier work on preemption-based prioritization [12], our approach explicitly models admission control with guaranteed resource allocation. The findings are consistent with survey studies on slicing performance [4, 5, 7], while extending them by quantifying blocking and interruption trade-offs.

From a practical perspective, the model illustrates how operators can tune Q_1 and Q_2 to balance eMBB and URLLC requirements. Such tuning directly impacts blocking probabilities and guarantees acceptable QoS under varying load conditions. This insight is useful for SLA negotiation and capacity planning.

Naturally, the model has limitations. It assumes Poisson arrivals, exponential service times, and only two slices. Correlations between traffic types, mobility patterns, or more complex slice hierarchies are not captured. Moreover, the interruption mechanism is simplified, focusing only on guaranteed thresholds.

Future extensions could include machine-learning-based controllers for adaptive RAC [18, 19], as well as integration with edge computing and 6G scenarios [16, 20].

7. Conclusions

This paper introduced a radio admission control model for 5G network slicing with priority interruption. The model differentiates between streaming and elastic traffic and evaluates key performance metrics such as blocking, accepting, and interruption probabilities, as well as mean numbers of active requests.

The results demonstrate that the proposed mechanism successfully captures the trade-offs between heterogeneous traffic classes. The analysis provides operators with quantitative guidelines for configuring guaranteed capacities and balancing competing QoS requirements. The presented framework thus offers a foundation for more efficient and fair resource allocation strategies in slice-aware wireless networks.

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

Funding: This research was funded by the Russian Science Foundation, project no. 24-19-00804 (<https://rscf.ru/en/project/24-19-00804/>).

Data Availability Statement: All data generated or analyzed during this study are included in this article. Additional numerical results and simulation scripts are available from the corresponding author upon reasonable request.

Acknowledgments: The authors thank the Department of Probability Theory and Cybersecurity of RUDN University for valuable discussions and methodological support.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Declaration on Generative AI: The authors have not employed any Generative AI tools.

References

- [1] J. Meredith, F. Firmin, and M. Pope, “Release 16 Description; Summary of Rel-16 Work Items,” 3rd Generation Partnership Project (3GPP), Tech. Rep., version 16.2.0, 2022.

- [2] M. Zambianco, A. Lieto, I. Malanchini, and G. Verticale, "A Learning Approach for Production-Aware 5G Slicing in Private Industrial Networks," in *Proceedings of the ICC 2022—IEEE International Conference on Communications*, Seoul, Korea, May 2022.
- [3] S. Elayoubi, M. Fallgren, P. Spapis, L. Campoy, B. Martinez, A. Mavromatis, and P. Iovanna, "5G Service Requirements and Operational Use Cases: Analysis and Implementation," *IEEE Communications Magazine*, vol. 54, no. 11, pp. 84–90, 2016. DOI: 10.1109/MCOM.2016.1600172CM
- [4] X. Foukas, G. Patounas, A. Elmokashfi, and M. Marina, "Network Slicing in 5G: Survey and Challenges," *IEEE Communications Magazine*, vol. 55, no. 5, pp. 94–100, 2017. DOI: 10.1109/MCOM.2017.1600951
- [5] H. Zhang, N. Liu, X. Chu, K. Long, A. Aghvami, and V. Leung, "Network Slicing Based 5G and Future Mobile Networks: Mobility, Resource Management, and Challenges," *IEEE Communications Magazine*, vol. 55, no. 8, pp. 138–145, 2017. DOI: 10.1109/MCOM.2017.1600940
- [6] M. Richart, J. Baliosian, J. Serrat, and J. Gorricho, "Resource slicing in virtual wireless networks: A survey," *IEEE Transactions on Network and Service Management*, vol. 13, no. 3, pp. 462–476, 2016. DOI: 10.1109/TNSM.2016.2597290
- [7] P. Rost, A. Banchs, I. Berberana, M. Breithach, M. Doll, H. Droste, C. Mannweiler, M. Puente, K. Samdanis, and B. Sayadi, "Network Slicing to Enable Scalability and Flexibility in 5G Mobile Networks," *IEEE Communications Magazine*, vol. 55, no. 5, pp. 72–79, 2017. DOI: 10.1109/MCOM.2017.1600939
- [8] V. Tikhvinskiy and G. Bochechka, "Prospects and QoS requirements in 5G networks," *Journal of Telecommunications and Information Technology*, vol. 2015, no. 1, pp. 23–26, 2015.
- [9] J. Ordonez-Lucena, P. Ameigeiras, D. Lopez, J. Ramos-Munoz, J. Lorca, and J. Folgueira, "Network Slicing for 5G with SDN/NFV: Concepts, Architectures, and Challenges," *IEEE Communications Magazine*, vol. 55, no. 5, pp. 80–87, 2017. DOI: 10.1109/MCOM.2017.1600935
- [10] F. Zhou, R. Hu, Z. Li, and Z. Han, "Network slicing as a service: enabling enterprises' own software-defined cellular networks," in *IEEE Communications Magazine*, vol. 54, 2016, pp. 146–153.
- [11] R. Ferrus, O. Sallent, J. Perez-Romero, and R. Agusti, "On 5G Radio Access Network Slicing: Radio Interface Protocol Features and Configuration," *IEEE Communications Magazine*, vol. 56, no. 5, pp. 184–192, 2018. DOI: 10.1109/MCOM.2018.1700659
- [12] Y. Adou, E. Markova, and Y. Gaidamaka, "Modeling and analyzing preemption-based service prioritization in 5G networks slicing framework," *Future Internet*, vol. 14, no. 10, p. 299, 2022. DOI: 10.3390/fi14100299
- [13] F. Kamoun and L. Kleinrock, "Analysis of shared finite storage in a computer networks node environment under general traffic conditions," *IEEE Transactions on Communications*, vol. 28, no. 7, pp. 992–1003, 1980. DOI: 10.1109/TCOM.1980.1094756
- [14] B. Han, V. Gopalakrishnan, L. Ji, and S. Lee, "Network Function Virtualization: Challenges and Opportunities for Innovations," *IEEE Communications Magazine*, vol. 53, no. 2, pp. 90–97, 2015. DOI: 10.1109/MCOM.2015.7045396
- [15] I. Parvez, A. Rahmati, I. Guvenc, A. Sarwat, and H. Dai, "A Survey on Low Latency Towards 5G: RAN, Core Network and Caching Solutions," *IEEE Communications Surveys & Tutorials*, vol. 20, no. 4, pp. 3098–3130, 2018. DOI: 10.1109/COMST.2018.2841349
- [16] A. Ateya, A. Alhussan, H. Abdallah, M. Al duailij, A. Khakimov, and A. Muthanna, "Edge Computing Platform with Efficient Migration Scheme for 5G/6G Networks," *Computer Systems Science and Engineering*, vol. 45, no. 2, pp. 1775–1787, 2023.
- [17] G. Association, "Generic Network Slice Template," GSM Association Official Document NG.116, Tech. Rep., version 1.0, May 2019.
- [18] D. Bega, M. Gramaglia, A. Banchs, X. Costa-Pérez, and V. Sciancalepore, "A Machine Learning Approach to 5G Infrastructure Market Optimization," *IEEE Transactions on Mobile Computing*, vol. 19, no. 3, pp. 498–512, 2019. DOI: 10.1109/TMC.2019.2893269
- [19] Y. Sun, M. Peng, Y. Zhou, Y. Huang, and S. Mao, "Application of Machine Learning in Wireless Networks: Key Techniques and Open Issues," *IEEE Communications Surveys & Tutorials*, vol. 21, no. 4, pp. 3072–3108, 2019. DOI: 10.1109/COMST.2019.2924243

- [20] S. Zhang, J. Xu, W. Wang, and X. Yang, “6G: Connectivity, Networking and Services,” *IEEE Internet of Things Journal*, vol. 8, no. 7, pp. 5399–5410, 2021. DOI: 10.1109/JIOT.2020.3037399

Information about the authors

Markova, Ekaterina V.—Docent, Ph.D. in Physics and Mathematics, Associate Professor of Department of Probability Theory and Cyber Security of RUDN University (e-mail: markova-ev@rudn.ru, ORCID: 0000-0002-7876-2801, Scopus Author ID: 56315828200)

Tazaeva, Anastasia A.—Student of Department of Probability Theory and Cyber Security of RUDN University, (e-mail: tazaeva081@gmail.com, ORCID: 0009-0001-4065-0847)

RUSSIAN METADATA

УДК 519.872, 519.217

DOI: 10.22363/2658-4670-2026-34-3-334-344

EDN: RRPVBC

Сценарий приоритетного обслуживания эластичного трафика в рамках концепции сетевой нарезки

А. А. Тазаева¹, Е. В. Маркова^{1,2}

¹ Российский университет дружбы народов, ул. Миклухо-Маклая, д. 6, Москва, 117198, Российская Федерация

² Институт проблем информатики, Федеральный исследовательский центр «Информатика и управление» РАН, ул. Вавилова, д. 44, корп. 2, Москва, 119333, Российская Федерация

Аннотация. *Предпосылки.* Технология сетевой нарезки (NS) позволяет создавать логические слои в рамках общей сетевой инфраструктуры, но сталкивается с проблемой совместного обслуживания разнородного трафика (потокowego и эластичного), предъявляющего различные требования к QoS. *Цель.* Разработать и провести анализ схемы управления доступом к радиоресурсам с приоритетным обслуживанием в сети 5G с двумя слоями, которая позволит обеспечить баланс между требованиями к обслуживанию потокowego и эластичного трафика. *Метод.* Используются методы теории массового обслуживания. Разработана двумерная марковская модель. Потоковой трафик предъявляет фиксированное требование к ресурсам; для эластичного трафика требование к ресурсам зависит от состояния системы. Реализован сценарий полного освобождения гарантированной эластичному трафику емкости. Распределение вероятностей состояний системы получено численно. *Результаты.* С ростом интенсивности поступления потокowego трафика вероятность прерывания передачи эластичного трафика сначала резко возрастает, затем стабилизируется. Среднее число обслуживаемых запросов стремится к гарантированным числам. Вероятность блокировки эластичного трафика почти не меняется. *Выводы.* Предложенная схема доступа к ресурсам сети представляет собой компромисс при обслуживании разных типов трафика. Модель позволяет получить количественные рекомендации для настройки гарантированных емкостей в системах 5G/6G.

Ключевые слова: 5G, нарезка сети, мобильные сети, QoS, услуги с гарантированным обслуживанием, услуги без гарантий к обслуживанию, показатели производительности, приоритетное обслуживание, прерывание обслуживания, управление доступом к ресурсам, распределение ресурсов