Network slice degradation probability as a metric for defining slice performance isolation*

Nikita Polyakov[1][0000−0003−0152−9646], Natalia Yarkina[2][0000−0003−3197−2737], Konstantin Samouylov[1,3][0000−0002−6368−9680], and Yevgeni Koucheryavy[2][0000−0003−3976−297X]

[1] RUDN University, Miklukho-Maklaya St. 6, Moscow 117198, Russia
[2] Unit of Electrical Engineering, Tampere University, 33720 Tampere, Finland
[3] Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, 44-2 Vavilov St, Moscow, 119333, Russia

Abstract. Slice isolation is a key feature of the network slicing technique and refers to protecting slices from negative impact of fault, attack or workload increase in other slices. Dynamic resource slicing policies, although provide efficient multiplexing and resource utilization, may lead to situation when a traffic surge in one slice hinders performance of other slices. The level of performance isolation cannot be specified and evaluated without defining an adequate metric. This paper addresses network slice degradation probability as a metric for defining performance isolation of slice. We use teletraffic theory to derive an analytical expression for the degradation probability in a single slice.

Keywords: Network slicing · Slice isolation · Slice degradation

1 Introduction

By logically decoupling the network, the infrastructure owner has the ability to dynamically reallocate virtualized network resources among slice tenants, which permits increasing both the performance and the commercial efficiency of the equipment. However, when applying the network slicing technique in the context of logical segmenting for virtual operators (slice tenants), the requirements for the isolation level and availability of network services must be satisfied [8]. According to the definition of virtual resources' isolation given in [4], a network slice has access to specific range of resources that do not overlap with other network slices.

The aim of this work is to define a metric for slice availability and isolation level while taking into account the dynamic and adaptive character of resource slicing. Reference [7] defines availability of an item as being in a state to perform a required function at a given instant of time or at any instant of time within a given time interval. Availability is usually expressed as a percentage of uptime

* Supported by the RUDN University Strategic Academic Leadership Program. The reported study was funded by RFBR, project number 19-07-00933, 20-07-01064.

in a day, month or year. According to [4], communication service availability is a percentage value of the amount of time the end-to-end communication service is delivered according to an agreed QoS, divided by the amount of time the system is expected to deliver the end-to-end service. We propose to define slice availability via service level degradation probability metric. Slice service/SLA degradation was mentioned in [2, 11], but not as a key concept.

The concept of service degradation reflects the specifics of packet-switched networks and describes a situation when, if a shortage of resources occurs, an incoming service request is not dropped or queued, but is satisfied, which leads to a temporary decrease in the amount of resources allocated to this and other ongoing sessions below the accepted value. The concept of degradation can be applied to both streaming and elastic traffic. In the first case, it will correspond, for example, to a temporary deterioration in the quality (resolution) of video during a video conference, while in the second case it will describe a breach of the transmission delay requirement.

This work proposes a probabilistic model of a slice which provides a communication service involving elastic (non-GBR) traffic transmission, such as software updates or buffered video streaming. We analyze the availability of the slice in terms of compliance with the SLA between the infrastructure owner and the tenant while maximizing economic benefits. We specifically focus on analytical modeling of slices with non-GBR services, while many authors, e.g. [8], derive formulas for GBR services or analyze system on equipment level [5, 1]. In this work, we consider the availability of slices only in terms of the compliance with QoS requirements, without taking into account equipment breakdowns and other external factors.

The paper is structured as follows. Section 2 introduces the basic model assumptions. Section 3 describes a probabilistic model in terms of queuing theory, for which formulas for performance indicators are further derived. Section 5 offers numerical results. Section 6 concludes the paper.

2 Basic Assumptions

For the system model, we overall follow [9] and [6]. We assume the total maximum transmission resource capacity of a 5G base station (BS), $C_{[\text{Gbps}]}$, constant and use it as the total amount of resources to be sliced. Let S slices be instantiated at the BS. For each slice s the following parameters are specified in the SLA:

- $R_s^{\min}$ — minimum data rate,
- $R_s^{\max}$ — maximum data rate,
- N_s^{cont} — contractual maximum number of users for which QoS will be guaranteed,
- R_s^{deg} — data rate threshold under which the slice is considered degraded.

The current state of slice s is assumed to be characterized by the number of active users, N_s, and the current capacity, C_s. The data rate for each slice user is then obtained as $R_s = C_s/N_s$.

By signing a service level agreement, the infrastructure owner commits to provide the tenant with a capacity (bandwidth) $C_s = N_s R_s^{\min}$ as long as $N_s \leq N_s^{\text{cont}}$, where $R_s^{\min}$ and N_s^{cont} are specified in the agreement. If $N_s > N_s^{\text{cont}}$ then only capacity $C_s^{\text{cont}} = N_s^{\text{cont}} R_s^{\min}$ is guaranteed. This bandwidth can be expressed as a share of the total resource capacity $\gamma_s = C_s^{\text{cont}}/C$. In other words, according to the SLA, slice users are guaranteed the data rate $R_s^{\min}$ as long as their number does not exceed N_s^{cont}, while any capacity above that can be provided only if available and without exceeding $R_s^{\max}$ per user.

When the slice bandwidth C_s is not enough to provide end users with the guaranteed minimum data rate $R_s^{\min}$, we talk about degradation of the service level in the slice. Degradation may occur due to an excessive workload increase in the slice itself (the tenant's responsibility) or due to insufficient resource allocation (the infrastructure owner's responsibility). Only in the latter case we refer to the slice isolation breach, i.e. whenever

$$C_s < R_s^{\min} \min\{N_s, N_s^{\text{cont}}\} \stackrel{\text{not}}{=} C_s^{\text{guar}}. \tag{1}$$

Thus, whenever the above condition is true, the availability of the slice is compromised. Moreover, the slice may lose availability due to some other reasons such as equipment failure. We assume that the required availability ratio Av is specified for each slice in the SLA. Reference [3] describes how to apply the reliability theory to the design of wireless networks and defines the steady-state availability as follows:

$$Av = \lim_{T\to\infty} Av(T), \tag{2}$$

where $Av(T) = \mathbb{P}\{\text{“equipment is up at time } T\text{”}\}$. Let us rewrite this as

$$Av = \lim_{T\to\infty} \frac{t^{\,\text{up}}(T)}{T} = 1 - \lim_{T\to\infty} \frac{t^{\,\text{fail}}(T) + t^{\,\text{cap}}(T)}{T}, \tag{3}$$

where T is the total measurement time, $t^{\,\text{up}}(T)$ is the total time of service up, $t^{\,\text{fail}}(T)$ is the total time of service failure and recovery and $t^{\,\text{cap}}(T)$ is the total time when the condition (1) holds.

In this work we do not take into account technical breakdowns, therefore we assume that $t^{\,\text{fail}}(T) = 0$.

Fig. 1 illustrates the periods $\Delta_{s,i}^{\text{cap}}$ when the slice capacity falls below the threshold. For simplicity, in the figure it is assumed that $N_s = N_s^{\text{cont}}$ at all times. Here

$$t_s^{\text{cap}}(T) = \sum_{i=1}^{Y_s(T)} \Delta_{s,i}^{\text{cap}}, \quad s = 1, ...S, \tag{4}$$

where $Y_s(t)$ is the counter of the periods when (1) holds.

3 Stochastic model of a single slice

Queuing theory can provide a convenient analytical framework to model the system described in the previous section. Each slice is modeled as a separate

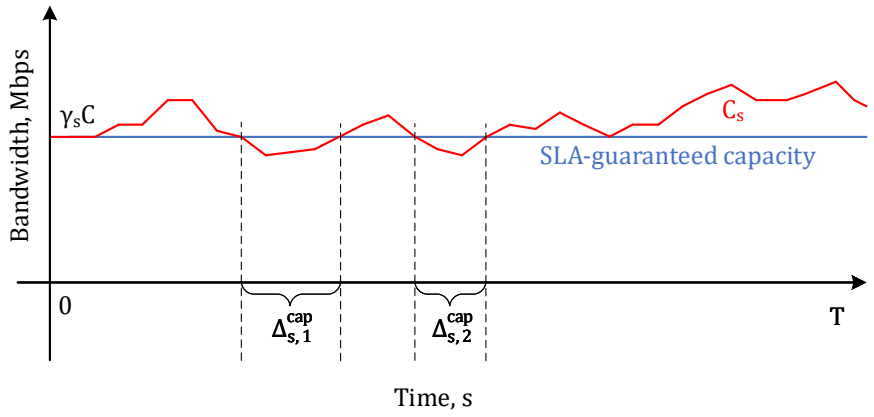


Fig. 1: Periods of slice s isolation breaches assuming that $N_s = N_s^{\text{cont}}$ for all T.

queuing system (QS), which type can be chosen so that it adequately reflects the nature of the service provided in the slice. Jobs in the QS correspond to user sessions. Since in the system model the slices are part of a single network with a total capacity of C, S queuing systems must share a common resource (capacity) C, and its part available to the QS s is denoted by C_s, $s = 1, ..., S$.

Here we consider only one slice type – the one providing best effort (BE) services with a data rate bounded above. Let us denote this type of slices by $BE^{\max}$. It is represented by a QS with egalitarian processor sharing (EPS) discipline with a total resource capacity of C (instead of 1). As in EPS [10], the service rate depends of the resource share per job, R_s, which is determined by their number. However, the resource share per job cannot exceed $R_s^{\max}$, which yields

$$R_s = \min\left\{\frac{C}{N_s}, R^{\max}\right\}. \tag{5}$$

Such a QS well reflects the provision of services related to the transfer of files or buffered streaming video.

Dynamic resource slicing in this context corresponds to the repeated redistribution of the capacity C among independent QSes according to some slicing policy and following the workloads.

The model under consideration is shown in Fig. 2, where $A_s(x)$ is the distribution law of intervals between session requests, and $B_s(x)$ is the distribution law of the job (file) sizes (corresponding to the service time on a resource unit) for $s = 1, ..., S$. It should be noted that admission control and resource allocation within a slice are individual characteristics for each slice type. For $BE^{\max}$, it is assumed that resources are shared equally among all users (jobs) and the number of jobs in service is unlimited. Since there is no admission control or queue in this slice type, there is a possibility that at a high arrival rate, the number

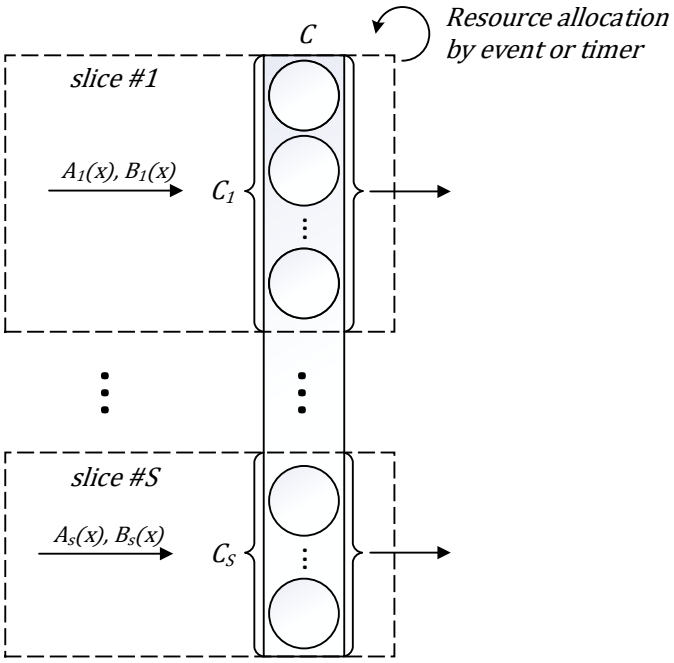


Fig. 2: System of S slices

of sessions will be so large that the user data rate will become unacceptable. Therefore, it makes sense to introduce a service speed degradation threshold, $R_s^{\deg}$, $0 < R_s^{\deg} \leqslant C$. The threshold will act as an indicator that the service is poorly provided. A decrease in the quality of service in a slice (degradation of a slice) can occur as a result of the arrival of the next request and/or re-slicing of the capacity.

In what follows, for convenience, the term "slice" will be used in the context of a probabilistic model and correspond to the term "QS".

Now let us consider the operation of a single $BE^{\max}$ slice in the system in Fig. 2. In this work we assume that slicing is called only upon model initialization and the slice's capacity, C_s is constant. For convenience, since we are considering one QS, until the end of this section we will omit the index s. Also, by the system we mean this QS.

Let the distribution laws $A(x) = Exp(\lambda)$ and $B(x) = Exp(\theta^{-1})$. Thus, the parameter λ is the mean number of session requests offered per time unit (the arrival rate), and θ is the average size of a job, since elastic traffic is considered. The maximum service speed, which corresponds to the amount of resources allocated per job, is bounded above by $R^{\max}$. The state space of such a QS can be divided into two parts: the states when jobs are served with the maximum speed $R^{\max}$, and when jobs receive less resources. Let M be the maximum number of

jobs in the system that can obtain the resource amount of $R^{\max}$,

$$M = \left\lfloor \frac{C}{R^{\max}} \right\rfloor. \tag{6}$$

Let us denote by $N(t)$ the number of jobs in the QS at time $t \geq 0$. The state space of the continuous-time Markov chain $N(t)$ has the form $\{0, 1, 2, ...\}$. The transition diagram of $N(t)$ is shown in Fig. 3. Once the number of jobs exceeds M, the service rate reaches its maximum $\frac{C}{\theta}$.

The stationary state probability distribution $\{p_n,\ n \geq 0\}$ exists if and only if

$$\lambda < \frac{C}{\theta}. \tag{7}$$

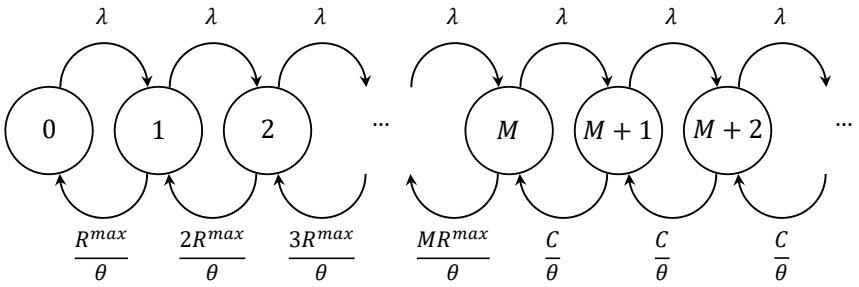


Fig. 3: Transition diagram

Let

$$\rho = \frac{\lambda}{\theta^{-1}} \tag{8}$$

Then the stationary state probabilities are given by

$$p_n = \begin{cases} \dfrac{1}{n!}\left(\dfrac{\rho}{R^{\max}}\right)^n p_0, & \text{if } 1 \leq n \leq M, \\ \dfrac{1}{M!}\dfrac{\rho^n}{C^{n-M}(R^{\max})^M} p_0, & \text{if } n \geq M. \end{cases} \tag{9}$$

or, in a more compact form,

$$p_n = \frac{1}{\min\{n,\ M\}!}\frac{\rho^n}{C^{\max\{0,\, n-M\}}(R^{\max})^{\min\{n,\, M\}}} p_0, \quad n \geq 1. \tag{10}$$

Let us find p_0 from the normalization condition :

$$\sum_{n=0}^{\infty} p_n = 1 \implies p_0 = \left(1 + \sum_{n=1}^{M-1} \frac{1}{n!}\left(\frac{\rho}{R^{\max}}\right)^n + \right.$$
$$\left. + \frac{1}{M!}\left(\frac{\rho}{R^{\max}}\right)^M \sum_{n=M}^{\infty}\left(\frac{\rho}{C}\right)^{n-M}\right)^{-1} = \tag{11}$$
$$= \left(1 + \sum_{n=1}^{M-1} \frac{1}{n!}\left(\frac{\rho}{R^{\max}}\right)^n + \frac{1}{M!}\left(\frac{\rho}{R^{\max}}\right)^M \frac{C}{C-\rho}\right)^{-1}.$$

provided that $\left|\frac{\lambda\theta}{C}\right| < 1$.

4 Slice Performance Measures

The stationary state probabilities obtained in the previous section permit deriving several important performance measures of a $BE^{\max}$ slice. Thus, the slice degradation probability, $P^{\deg}$, can be expressed as the probability that the number of jobs in the system exceeds $\lfloor C/R^{\deg} \rfloor$:

$$P^{\deg} = 1 - \sum_{n=0}^{\lfloor C/R^{\deg} \rfloor} p_n. \tag{12}$$

The required availability ratio Av, as we defined, is equipment uptime, therefore (excluding other factors):

$$Av = 1 - \mathbb{P}\left\{\frac{C}{R^{\min}} < \min\{N, N^{\text{cont}}\}\right\}. \tag{13}$$

Let UTIL denote the average system utilization. If $0 \le n \le M$, then the system resource C is not fully used, but only its part equal to $nR^{\max}$ is allocated. In other cases, when $n > M$, the entire resource is utilized. Therefore,

$$\text{UTIL} = \frac{\sum_{n=0}^{M} nR^{\max}p_n + \sum_{n=M+1}^{\infty} Cp_n}{C} =$$
$$= \frac{R^{\max}}{C}\sum_{n=0}^{M} np_n + \frac{p_0}{M!}\left(\frac{\rho}{R^{\max}}\right)^M \frac{\rho}{C-\rho}. \tag{14}$$

or, by replacing $\sum_{n=M+1}^{\infty} p_n$ with $1 - \sum_{n=0}^{M} p_n$,

$$\text{UTIL} = 1 - \sum_{n=0}^{M}(1 - \frac{nR^{\max}}{C})p_n. \tag{15}$$

The average number of jobs in service, N^{avg}, can be found as

$$N^{\text{avg}} = \sum_{n=1}^{\infty} np_n = \\ = \sum_{n=1}^{M-1} np_n + Cp_M \left(\frac{M}{C-\rho} + \frac{\rho}{(C-\rho)^2} \right). \tag{16}$$

The average service speed, or the amount or resources per user, is given by

$$R^{\text{avg}} = R^{\max} \sum_{n=1}^{M} p_n + \sum_{n=M+1}^{\infty} \frac{C}{n} p_n = R^{\max} \sum_{n=1}^{M} p_n - \\ - Cp_M \left(\frac{C}{\rho} \right)^M \left(\ln(1 - \frac{\rho}{C}) + \sum_{n=1}^{M} \frac{1}{n} \left(\frac{\rho}{C} \right)^n \right). \tag{17}$$

Finally, the average service time is determined according to Little's law:

$$T^{\text{avg}} = \frac{N^{\text{avg}}}{\lambda}. \tag{18}$$

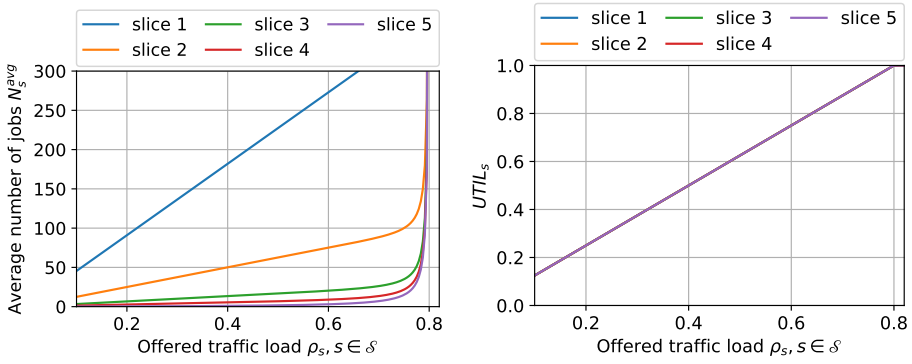
Table 1: Parameter values

Slice #	1	2	3	4	5
Traffic type	Buffered Video Streaming				Files Download
Description of traffic	SD	HD	UHD (4K)	VR (8K)	Software Update
$R_s^{\min}$, Mbps	2	5	25	50	30
$R_s^{\max}$, Mbps	2.2	8	30	75	C
Average file size θ, GB	0.3	1.2	2.5	5	1

5 Numerical Results

As an example, let us consider five slices with the parameters given in Table 1. We suppose that the five slices are allocated $C = 4$ Gbps and this capacity is shared among them equally. Traffic in the slices has a different degree of elasticity, increasing from slice 1 to 5. The aim of the numerical experiment here is to compare the behavior of the performance measures of slices that have different traffic characteristics. In Fig. 4–5 selected performance measures are plotted vs. the offered load ρ_s, while in Fig. 6–7 we additionally take into account the difference of the minimum user data rate $R_s^{\min}$.

Fig. 4b shows that the capacity utilization grows to its maximum value of 1 when the offered traffic ρ reaches a value close to 0.8 for all slices. In this case,



(a) (b)

Fig. 4: The average number of jobs N_s (a) and the capacity utilization (b) vs. the offered traffic $\rho_s = \rho$.

the average number of jobs N_s^{avg} in a slice grows hyperbolically (Fig. 4a). The more elastic the traffic in the slice, the steeper is the growth pattern.

Fig. 5a shows the conditional average user data rate given that the slice is not empty:

$$R_s^{\text{avg}*} = \frac{R_s^{\text{avg}}}{1 - p_0}. \tag{19}$$

This indicator also depends on the nature of the traffic.

The average job service time T_s shown in Fig. 5b demonstrates a different dependency: for higher values of ρ, sessions in slice 5 are longer than in slice 4, which has a higher data rate minimum $R_s^{\min}$, whereas when ρ is low, on the contrary, slice 5 having no constraint on the maximum data rate $R_s^{\max}$ shows the best result.

Now consider the degradation probability P_s^{deg} in slices in the range from 0 to 10%. The behavior of this metric in this case is difficult to relate to other indicators, so let us add a dependency on the minimum data rate $R_s^{\min}$ to the plots in Fig. 6–7.

It can be observed in the figures that With a uniform increase in the value of $\rho_s R_s^{\min}$:

- the average number of jobs in a slice still grows hyperbolically (Fig. 6a), but slice 5 overflows faster;
- the capacity utilization in Fig. 6b now behaves differently confirming that slice 5 is overflowing faster;
- the average user data rate (Fig. 7a) and job service time (Fig. 7b) additionally reflect the influence of the threshold of the maximum user data rate $R_s^{\max}$.

The degradation probability P_s^{deg} as a function of $\rho_s R_s^{\min}$ shows the cumulative effect of slice characteristics:

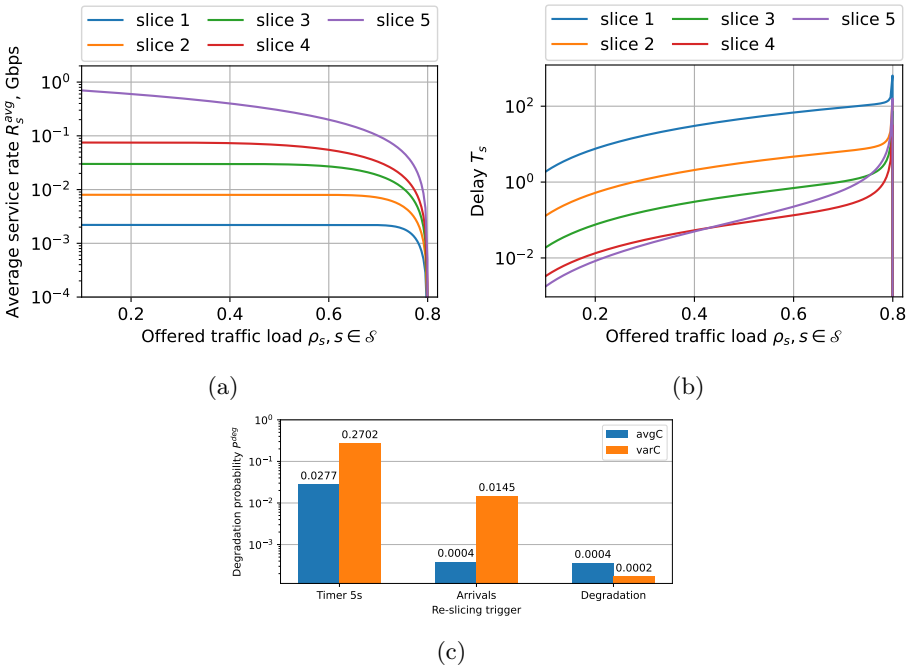
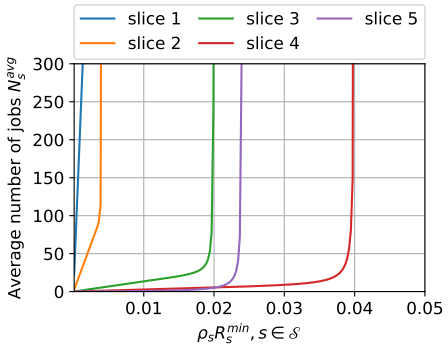


Fig. 5: The average user data rate $R_s^{\text{avg}*}$ (a), the average job service time T_s (b) and the degradation probability P_s^{deg} (c) vs. the offered traffic $\rho_s = \rho$.

– Slices 1 and 2 with the least elastic traffic and the smallest job sizes go into degradation much faster. Such traffic is more sensitive to increased load on the system.
– Slice 5 with the most elastic traffic degrades faster than slices 3, 4. This suggests that in general the ratio of the offered traffic load ρ to the threshold of the minimum user data rate $R_s^{\min}$ is of greater importance for the quality of service provided under a static resource slicing, than the threshold of the maximum user data rate $R_s^{\max}$.

6 Conclusion

In this paper we take a closer look at the important metrics for the application of 5G network slicing policies – slice availability and slice degradation probability. We provide an analytical model for a single best-effort slice with a bounded above user data rate. For this model, analytical formulas are derived to calculate the probability of slice degradation and other performance measures.



(a) (b)

Fig. 6: The average number of jobs (a) and the capacity utilization (b) vs. the offered traffic $\rho_s R_s^{\min}$.

References

1. Ateya, A.A., Muthanna, A., Gudkova, I., Vybornova, A., Koucheryavy, A.: Intelligent core network for tactile internet system. In: Proceedings of the International Conference on Future Networks and Distributed Systems. ICFNDS '17, Association for Computing Machinery, New York, NY, USA (2017). https://doi.org/10.1145/3102304.3102326, https://doi.org/10.1145/3102304.3102326
2. Brik, B., Ksentini, A.: On predicting service-oriented network slices performances in 5g: A federated learning approach. In: 2020 IEEE 45th Conference on Local Computer Networks (LCN). pp. 164–171 (2020). https://doi.org/10.1109/LCN48667.2020.9314849
3. Hößler, T., Scheuvens, L., Franchi, N., Simsek, M., Fettweis, G.P.: Applying reliability theory for future wireless communication networks. In: 2017 IEEE 28th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC). pp. 1–7 (2017). https://doi.org/10.1109/PIMRC.2017.8292773
4. NG, G.: 116–generic network slice template–version 5.0. Tech. rep., Technical report, 06 2021
5. Ometov, A., Kozyrev, D., Rykov, V., Andreev, S., Gaidamaka, Y., Koucheryavy, Y.: Reliability-centric analysis of offloaded computation in cooperative wearable applications. Wireless Communications and Mobile Computing **2017** (2017)
6. Polyakov, N.A., Yarkina, N.V., Samouylov, K.E.: A simulator for analyzing a network slicing policy with SLA-based performance isolation of slices. Discrete and Continuous Models and Applied Computational Science **29**(1), 36–52 (2021). https://doi.org/10.22363/2658-4670-2021-29-1-36-52
7. Recommendation, I.: E. 800: Terms and definitions related to quality of service and network performance including dependability (aug 1994)[cited 2013-11-20]. URL http://www. itu. int/rec/T-REC-E pp. 800–200809
8. Vilà, I., Pérez-Romero, J., Sallent, O., Umbert, A.: Characterization of radio access network slicing scenarios with 5g qos provisioning. IEEE Access **8**, 51414–51430 (2020). https://doi.org/10.1109/ACCESS.2020.2980685

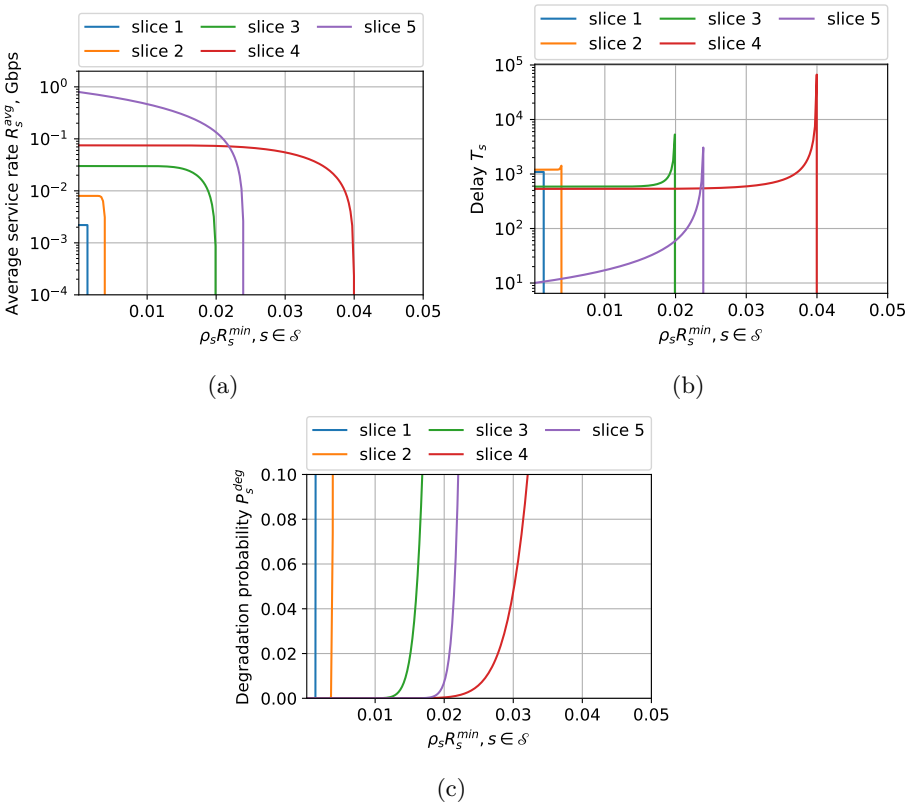


Fig. 7: The average user data rate (a), session duration (b) and degradation probability (c) vs. $\rho_s R_s^{\min}$.

9. Yarkina, N., Gaidamaka, Y., Correia, L., Samouylov, K.: An analytical model for 5G network resource sharing with flexible SLA-oriented slice isolation. Mathematics **8**, 1177 (07 2020). https://doi.org/10.3390/math8071177
10. Yashkov, S., Yashkova, A.: Processor sharing: A survey of the mathematical theory. Automation and Remote Control **68**(9), 1662–1731 (2007)
11. Zanzi, L., Sciancalepore, V.: On guaranteeing end-to-end network slice latency constraints in 5g networks. In: 2018 15th International Symposium on Wireless Communication Systems (ISWCS). pp. 1–6 (2018). https://doi.org/10.1109/ISWCS.2018.8491249