future internet

MDPI

Article

Streaming and Elastic Traffic Service in 5G-Sliced Wireless Networks and Mutual Utilization of Guaranteed Resource Units

Yves Adou[1,*], Ekaterina Markova [1,2,*] and Yuliya Gaidamaka [1,2]

[1] Institute of Computer Science and Telecommunications, Peoples' Friendship University of Russia Named after Patrice Lumumba (RUDN University), 6 Miklukho-Maklaya St, 117198 Moscow, Russia; gaydamaka-yuv@rudn.ru

[2] Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, 44-2 Vavilov St, 119333 Moscow, Russia

* Correspondence: akyb.access@gmail.com (Y.A.); markova-ev@rudn.ru (E.M.)

Abstract: Researchers of 5G-sliced Wireless Networks are faced with the task of achieving, at the same time, (i) mandatory isolation among network slices and (ii) the effective utilization of resource at Fifth Generation Base Station (gNB). This article proposes the second version of the Preemption-based service Prioritization (PP2) scheme that merges the capabilities of the classical Resource Reservation (RR) and Service Prioritization (SP) schemes and realizes the coexistence of non-TCP streaming and TCP elastic traffic at gNB. Analytical methods are given for Dynamic Resource Sharing (DRS) and Dynamic Resource Reallocation (DRR) with data rate reduction and service preemption. As main results, under some baseline scenario, the PP2 scheme, in comparison with RR, does increase the admission probability up to 99.999% at the system while ensuring 100% in system capacity utilization.

Keywords: 5G; slicing; quality of service; radio admission control; resource allocation; TCP elastic traffic; non-TCP streaming traffic

check for updates

Citation: Adou, Y.; Markova, E.; Gaidamaka, Y. Streaming and Elastic Traffic Service in 5G-Sliced Wireless Networks and Mutual Utilization of Guaranteed Resource Units. *Future Internet* **2024**, *16*, 397. https://doi.org/10.3390/fi16110397

Academic Editor: Jose I. Moreno Novella

Received: 24 August 2024
Revised: 7 October 2024
Accepted: 11 October 2024
Published: 30 October 2024

Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Four years after the Network Slicing (NS) official launch alongside other technologies (by far not the least important creations for Fifth Generation Base Stations (gNBs) [1–8]), major advances in the wireless telecommunication industry have been achieved with NS at the core, harboring features such as the inherently flexible Resource Reservation (RR) and/or Service Prioritization (SP) access scheme [1,2,4,6,8–14]. Unfortunately, these schemes do not completely enable NS's capabilities at the gNB: the RR proves especially inefficient during network outages, and the SP does not ensure NS's utmost important property—the slices' effective isolation. For the review, we refer the reader to [10].

The NS concept, standardized by Third Generation Partnership Project (3GPP) and implemented by Docomo and Ericsson, aims at efficient resource utilization while meeting the slices' aggregated Quality of Service (QoS) requirements [1,3–7]. To achieve its goal, NS first needs to overcome the above-mentioned RR and SP inefficiencies. An obvious solution, and by far the simplest, seems to be to just interconnect the RR isolation capability and SP efficient resource utilization in a Preemption-based service Prioritization (PP) scheme, so-called by research [10]. The setback at this point is that the PP scheme performs the expected job exclusively among slices, with each slice providing customers with non-TCP streaming service at a gNB.

The current research aimed to develop the second version of the PP (PP2) scheme—an improvement or adaptation of the PP scheme in order to allow the simultaneous provision of non-TCP streaming and TCP elastic services at a gNB. Thus, the remainder of this paper is outlined as follows. First, the system model, together with the optimization problem and solution, is presented in Section 2. Next, the mathematical model is constructed, and the relevant system Key Performance Indicators (KPIs) are described in Section 3. Then,

an accurate comparison of the PP2 scheme with the widely utilized RR is conducted in Section 4. Finally, this paper is concluded in Section 5.

2. System Model

Let us consider the coexistence of multiple slices at one Fifth Generation Base Station (gNB), each providing either non-TCP streaming or TCP elastic service (non-TCP streaming or TCP elastic slice) to customers (Figure 1). Thus, we consider a system model with a «service-type-based slice» [3,4,9]. We denote the following notation:
$C, 0 < C < \infty$—the total system capacity;
$\mathcal{K} = \{1, \ldots, K\}$—the finite set of slices in the system;
$\mathcal{K}_s \subseteq \mathcal{K}, |\mathcal{K}_s| = K_s$—the subset of non-TCP streaming slices;
$\mathcal{K}_e \subseteq \mathcal{K}, \mathcal{K}_s \cap \mathcal{K}_e = \emptyset, \mathcal{K}_s \cup \mathcal{K}_e = \mathcal{K}, |\mathcal{K}_e| = K_e$—the subset of TCP elastic slices;
$C_k, 0 < C_k \leq C, \sum_{i=1}^{K} C_i \geq C, k \in \mathcal{K}$—the overall capacity of slice k;
$Q_k, 0 \leq Q_k \leq C_k, \sum_{i=1}^{K} Q_i \leq C$—the guaranteed capacity of slice k.

We note that the model is built at the top level of abstraction, without particular attention paid to the configuration details at the lower layers of the OSI/ISO model. Thus, we consider the bitrate (one bit per second) as the resource unit (r.u.) since the Quality of Service (QoS) requirements for services are primarily specified in terms of bitrate. From this point of view, the slice capacity is measured in r.u.

Let the customer flow at a slice $k \in \mathcal{K}$ be a Poisson flow with intensity $\lambda_k \left[\frac{\text{customers}}{\text{time units (t.u.)}}\right]$. We assume that a customer flow $k_s \in \mathcal{K}_s$ is characterized by the service time with intensity $\mu_{k_s} \left[\frac{\text{customers}}{\text{t.u.}}\right]$ and constant data rate $b_{k_s} \left[\frac{\text{r.u.}}{\text{customer}}\right]$, while a customer flow $k_e \in \mathcal{K}_e$ denotes an exponentially distributed random size of the transmitted data block with parameter $\mu_{k_e}^{-1}$[size units (s.u.)] and minimum data rate $b_{k_e} \left[\frac{\text{r.u.}}{\text{customer}}\right]$. Let us note that r.u. are expressed in s.u. per t.u. $\left(\frac{\text{s.u.}}{\text{t.u.}}\right)$. Table 1 groups the characteristics of the flows in the system.

Table 1. The characteristics of a customer flow $k \in \mathcal{K}$.

Flow k	$\lambda_k, 0 < \lambda_k < \infty$	$\mu_k^{-1}, 0 < \mu_k^{-1} < \infty$	$b_k, 0 < b_k \leq C_k$	$\rho_k, 0 < \rho_k < \infty$
$k \in \mathcal{K}_s$ (Streaming service)	Intensity of the customers' arrival	Mean service time of one customer	Constant data rate of one customer	Offered load $\lambda_k \mu_k^{-1}$
$k \in \mathcal{K}_e$ (Elastic service)	Intensity of the data blocks' arrival	Mean size of one data block	Minimum data rate of one customer	Offered load $\lambda_k \mu_k^{-1}$

Let us denote the following:
$D_k = \left\lfloor C b_k^{-1} \right\rfloor, 0 < D_k < \infty, k \in \mathcal{K}$—the maximum number of customers in slice k, given that equality $C_k = C$ holds;
$M_k = \left\lfloor C_k b_k^{-1} \right\rfloor, 0 < M_k \leq D_k$—the maximum number of customers in slice k;
$N_k = \left\lfloor Q_k b_k^{-1} \right\rfloor, 0 < N_k \leq M_k$—the guaranteed number of customers in slice k;
$n_k, 0 \leq n_k \leq M_k$—the number of customers in slice k;
$\mathbf{n} := (n_1, \ldots, n_K)^T, \mathbf{n} \in \Omega$—the gNB state, with the state space

$$\Omega = \left\{ \mathbf{n} \in (\mathbb{N} \cup \{0\})^K : \sum_{k=1}^{K} n_k b_k \leq C \right\}, \quad (z)^+ = \begin{cases} z, & z > 0; \\ 0, & z \leq 0; \end{cases} \tag{1}$$

$c_k(\mathbf{n}), 0 \leq c_k(\mathbf{n}) \leq C_k$—the capacity share of slice k;
$x_k(\mathbf{n}) = n_k^{-1} c_k(\mathbf{n}), n_k > 0$—the data rate allocated to each customer in slice k (with this, we assume the uniform allocation of the capacity share of slice k).

According to [15], the state space $\Omega = \Omega^{\max} \cup \Omega^{\text{opt}}$, where $\Omega^{\max}$ denotes the feasible states that all customers can be allocated the maximum data rates:

$$\Omega^{\max} = \left\{ \mathbf{n} \in \Omega : \sum_{k_s \in \mathcal{K}_s} n_{k_s} b_{k_s} + \sum_{k_e \in \mathcal{K}_e} n_{k_e} C_{k_e} \leq C \right\}; \quad (2)$$

and $\Omega^{\text{opt}} = \Omega \setminus \Omega^{\max}$ denotes all other feasible states.

We note that two essential slicing features are illustrated in the states of Ω^{opt}:

- The total system capacity is occupied, which ensures that an r.u. is not idle if it can be used to provide services with the highest quality;
- The optimization of r.u. utilization can be performed while ensuring slice isolation for guaranteed QoS.

We denote $\mathcal{K}_v(\mathbf{n}) \subseteq \mathcal{K}, \mathbf{n} \in \Omega$ as the subset of slices violating the mandatory isolation among the slices (slice violators) in an arbitrary system state $\mathbf{n}$. Thus, a slice violator is a slice (non-TCP streaming or TCP elastic slice) that uses r.u. guaranteed to another slice (slice owner—also non-TCP streaming or TCP elastic slice) during time intervals when those r.u. are not required/claimed by the other slice due to a lack of customers.

We assume that the guaranteed capacity of a slice $k_o \in \mathcal{K}$ (slice owner) could be utilized by slices violators $k_v \in \mathcal{K}_v(\mathbf{n})$ in arbitrary system state $\mathbf{n}$. With this feature, QoS management and isolation among slices can be a problem in the system. Therefore, we propose an analytical solution method—the Dynamic Resource Reallocation (DRR). Through capacity share reduction, DRR can preempt customers' service in slices violators k_v, whenever required for a customer admission to slice owner k_o. We assume that DRR starts its job from slice violator k_v with the lowest Priority Level (PL) and keeps up until the required capacity is gathered [10,11]. To gather capacity shares of non-TCP streaming slices violators ($k_v \in \mathcal{K}_s$), DRR utilizes service preemption [10], while for TCP elastic slices violators ($k_v \in \mathcal{K}_e$), it does so after reducing the data rate allocated to each customer in TCP elastic slices violators k_v down to the minimum [11,15]. The DRR method is analytically described in Definition 2.

A depiction of the considered K-dimensional Queueing System (QS) is given by the scheme model in Figure 1.

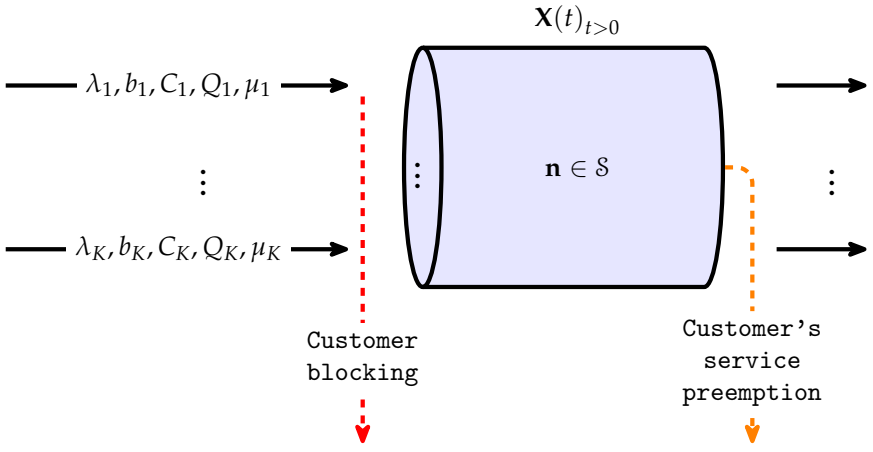


Figure 1. K-dimensional QS with customer blocking and service preemption.

Let us assume that in arbitrary system state $\mathbf{n} \in \Omega$, $\mathcal{K}^+(\mathbf{n}) = \{k \in \mathcal{K} : n_k > 0\}$ and $\mathcal{K}_e^+(\mathbf{n}) = \{k_e \in \mathcal{K}_e : n_{k_e} > 0\}$.

Note that each non-TCP streaming slice can be seen as 'TCP elastic-friendly' [16], meaning that the minimum data rate allocated to one customer equals the maximum, which makes it constant. Thus, following [11], we consider the next optimization problem with a utility function of the log type that was proposed by Kelly [17] for proportionally fair

Dynamic Resource Sharing (DRS), which, in our case, coincides with max–min fairness ([18], pp. 111–204):

$$\sum_{k\in\mathcal{K}^+(\mathbf{n})} W_k(n_k)n_k\ln(x_k) \longrightarrow \max, \tag{3}$$

$$\text{subject to} \sum_{k\in\mathcal{K}^+(\mathbf{n})} x_k n_k = C, \tag{4}$$

$$\text{over } \mathbf{x}\in\mathcal{P} = \left\{\mathbf{x}=(x_1,\ldots,x_K)^T : \bigwedge_{k\in\mathcal{K}^+(\mathbf{n})} (b_k\le x_k\le C_k)\right\}, \tag{5}$$

where $W_k(n_k) = \begin{cases} 1, & n_k\le N_k; \\ n_k^{-1}N_k, & n_k > N_k. \end{cases}$, $k\in\mathcal{K}$, denotes the weight functions [15].

We propose an analytical DRS solution method, allocating a data rate to each customer in the considered system.

Definition 1. *The DRS method is described by the following recursive sequence of numbers:*

$$w_k^{(0)}(\mathbf{n}) = \begin{cases} n_k b_k, & k\in\mathcal{K}_\mathrm{s}, \\ \min(n_k b_k, Q_k), & k\in\mathcal{K}_\mathrm{e}, \end{cases} \qquad k=\overline{1,K}, \tag{6}$$

$$w_k^{(l)}(\mathbf{n}) = \begin{cases} w_k^{(l-1)}(\mathbf{n})r_k^{(l-1)}(\mathbf{n}), & k\in\mathcal{K}_\mathrm{s}, \\ \min\left(C_k, w_k^{(l-1)}(\mathbf{n}) + \left(C-\sum_{i=1}^K w_i^{(l-1)}(\mathbf{n})\right)^+ r_k^{(l-1)}(\mathbf{n})\right), & k\in\mathcal{K}_\mathrm{e}, \end{cases} \qquad l=\overline{1,K_\mathrm{e}+1}, \tag{7}$$

where

$$r_k^{(0)}(\mathbf{n}) = \begin{cases} 1, & k\in\mathcal{K}_\mathrm{s}, \\ \left(\frac{n_k \mathrm{I}\left(\sum_{i=1}^K (n_i-N_i)^+=0\right)}{\sum_{j\in\mathcal{K}_\mathrm{e}} n_j} + \frac{(n_k-N_k)^+\mathrm{I}\left(\sum_{i=1}^K (n_i-N_i)^+>0\right)}{\sum_{j\in\mathcal{K}_\mathrm{e}}\left(n_j-N_j\right)^+}\right)^+, & k\in\mathcal{K}_\mathrm{e}, \end{cases} \tag{8}$$

$$r_k^{(l)}(\mathbf{n}) = \begin{cases} 1, & k\in\mathcal{K}_\mathrm{s}, \\ \left(\frac{n_k}{\sum_{j\in\mathcal{K}_\mathrm{e}} n_j \mathrm{I}\left(C_j-w_j^{(l)}(\mathbf{n})>0\right)}\right)^+, & k\in\mathcal{K}_\mathrm{e}, \end{cases} \tag{9}$$

given $\mathrm{I}(\textit{event } A) = \begin{cases} 1, & \textit{if } A \textit{ occurs}, \\ 0, & \textit{otherwise}, \end{cases}$ *, which is the indicator function. Let us denote the following:*
$c_k(\mathbf{n}) := w_k^{(K_\mathrm{e}+1)}(\mathbf{n})$*—the capacity share of slice k in arbitrary system state* $\mathbf{n}\in\Omega$*;*
$\sum_{k=1}^K c_k(\mathbf{n}) = c_1(\mathbf{n})+\ldots+c_K(\mathbf{n})$*—the allocated capacity of the system;*
$C-\sum_{k=1}^K c_k(\mathbf{n})$*—the unallocated capacity of the system.*

Explanations for Definition 1 are given by the illustrations in Figure 2.

Let us denote $\mathbf{e}_k, k\in\mathcal{K}$ as the k-th row of a $K\times K$ identity matrix.

Given the above DRS method, we organize the Radio Admission Control (RAC) scheme in such a way that, upon a customer's arrival at a slice $k\in\mathcal{K}$ in arbitrary system state $\mathbf{n}\in\Omega$, three paths are available:

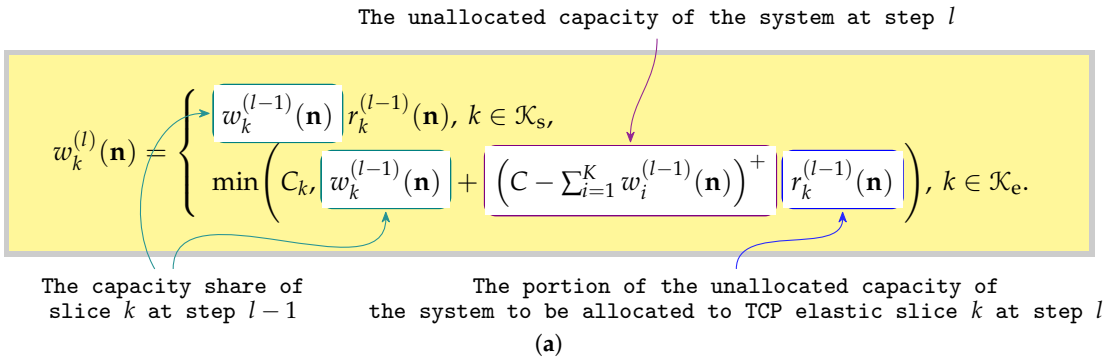
1. Admission via preemption—when the data rate allocated to each customer in any of the TCP elastic slices will be less than the minimum, and the number of customers in the slice k is less than guaranteed.
2. Blocking (or rejection) in two cases:
   (a) When the data rate allocated to each customer in any of the TCP elastic slices will be less than the minimum, and the number of customers in the slice k will be more than that guaranteed.
   (b) When the number of customers in the slice k is equal to the maximum.

3. Direct admission (no preemption)—when the data rate allocated to each customer in each TCP elastic slice will be equal or more than the minimum, and the number of customers in the slice k is less than the maximum.

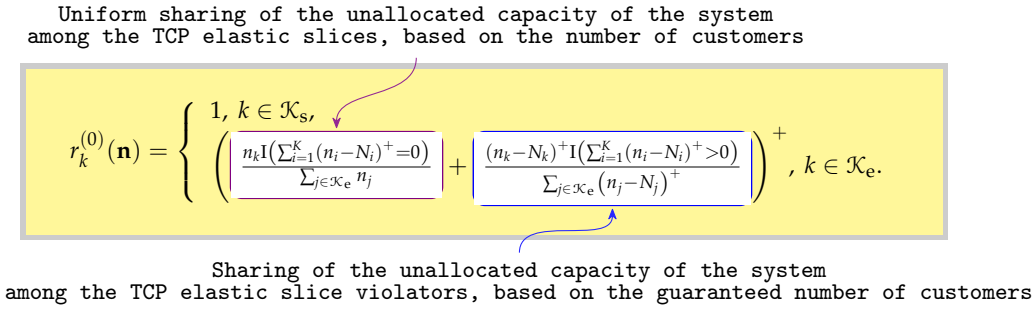
A depiction of the RAC scheme for accessing a slice $k \in \mathcal{K}$ is given in Figure 3.

Let us assume that in arbitrary system state $\mathbf{n}$, $\mathcal{K}_{\mathrm{v}}(\mathbf{n}) = \{k_{\mathrm{v}} \in \mathcal{K} : n_{k_{\mathrm{v}}} > N_{k_{\mathrm{v}}}\}$.

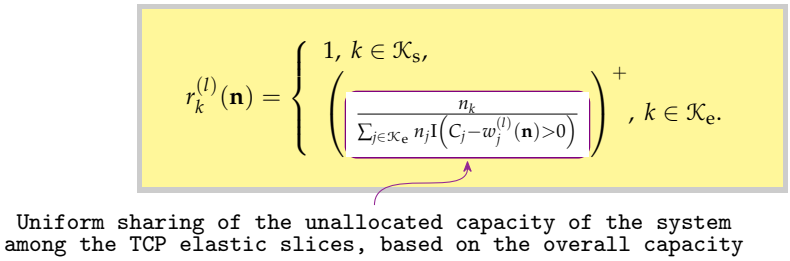
We analytically describe the DRR method governing the preemption for a customer admission to a slice owner $k_{\mathrm{o}} \in \mathcal{K}$ in arbitrary system state $\mathbf{n} \in \Omega$. We denote the following:
$u_{k_{\mathrm{o}}}^{k_{\mathrm{v}}}(\mathbf{n}), k_{\mathrm{v}} \in \mathcal{K}_{\mathrm{v}}(\mathbf{n})$—the capacity share of slice violator k_{v} to reallocate for a customer admission to slice owner k_{o};
$m_{k_{\mathrm{o}}}^{k_{\mathrm{v}}}(\mathbf{n})$—the corresponding number of customers to preempt.



(a)



(b)



(c)

Figure 2. Explanations for Definition 1 (Dynamic Resource Sharing (DRS)). (**a**) Details about Equation (7); (**b**) Details about Equation (8); (**c**) Details about Equation (9).

Definition 2. *The DRR method is based on the following sequence of numbers in arbitrary system state* $\mathbf{n} \in \Omega$*:*

$$u_{k_{\mathrm{o}}}^{K}(\mathbf{n}) = \min\left((c_K(\mathbf{n}) - Q_K)^{+}, (\delta_{k_{\mathrm{o}}}(\mathbf{n}) - C)^{+}\right), \quad k_{\mathrm{o}} \in \mathcal{K}, \tag{10}$$

$$u_{k_{\mathrm{o}}}^{l}(\mathbf{n}) = \min\left((c_l(\mathbf{n}) - Q_l)^{+}, \left(\delta_{k_{\mathrm{o}}}(\mathbf{n}) - \sum_{j=l+1}^{K} u_{k_{\mathrm{o}}}^{j}(\mathbf{n}) - C\right)^{+}\right), \quad l = \overline{K-1, 1}, \tag{11}$$

with $\delta_{k_o}(\mathbf{n}) = \sum_{j=1}^{K} c_j(\mathbf{n}) + b_{k_o}$ *means that we require capacity* b_{k_o}. *Therefore, we obtain*

$$m_{k_o}^{k}(\mathbf{n}) = \left(n_k - \left\lfloor \frac{c_k(\mathbf{n}) - u_{k_o}^{k}(\mathbf{n})}{b_k} \right\rfloor \right)^{+}, \quad k \in \mathcal{K}. \tag{12}$$

Using the DRR method for a customer admission to a slice owner $k_o \in \mathcal{K}$ in an arbitrary system state $\mathbf{n} \in \Omega$, one may note the following:
$0 \le u_{k_o}^{k_v}(\mathbf{n}) \le C_{k_v} - Q_{k_v}$, $0 \le m_{k_o}^{k_v}(\mathbf{n}) \le M_{k_v} - N_{k_v}$, $k_v \in \mathcal{K}_v(\mathbf{n})$—the capacity share of slice violator k_v to reallocate cannot exceed $C_{k_v} - Q_{k_v}$, which means that the corresponding number of customers to preempt cannot exceed $M_{k_v} - N_{k_v}$;
$0 \le \sum_{j=1}^{K} u_{k_o}^{j}(\mathbf{n}) \le b_{k_o}$—the total capacity to reallocate cannot exceed the data rate b_{k_o}.

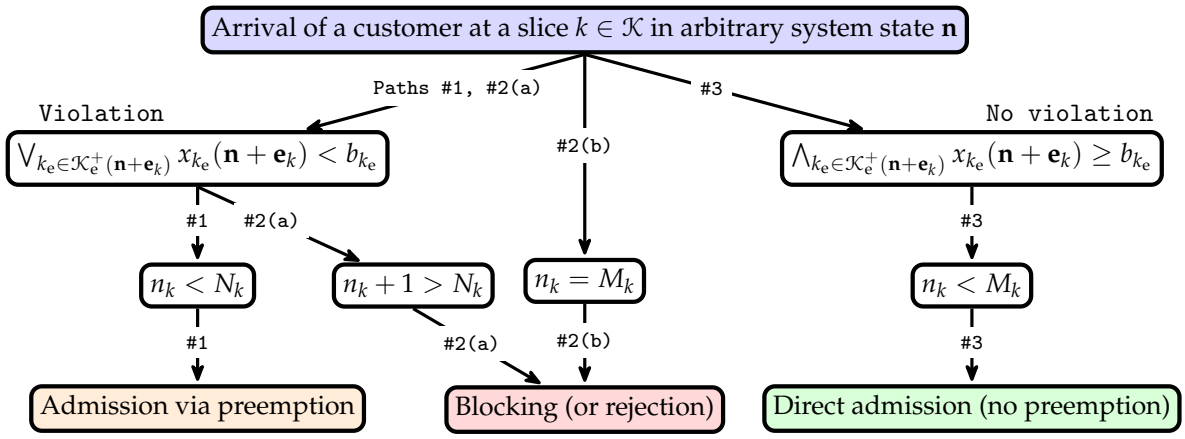


Figure 3. Formalization of the RAC scheme for accessing a slice $k \in \mathcal{K}$ in the system.

Explanations for Definition 2 are given by the illustrations in Figure 4.

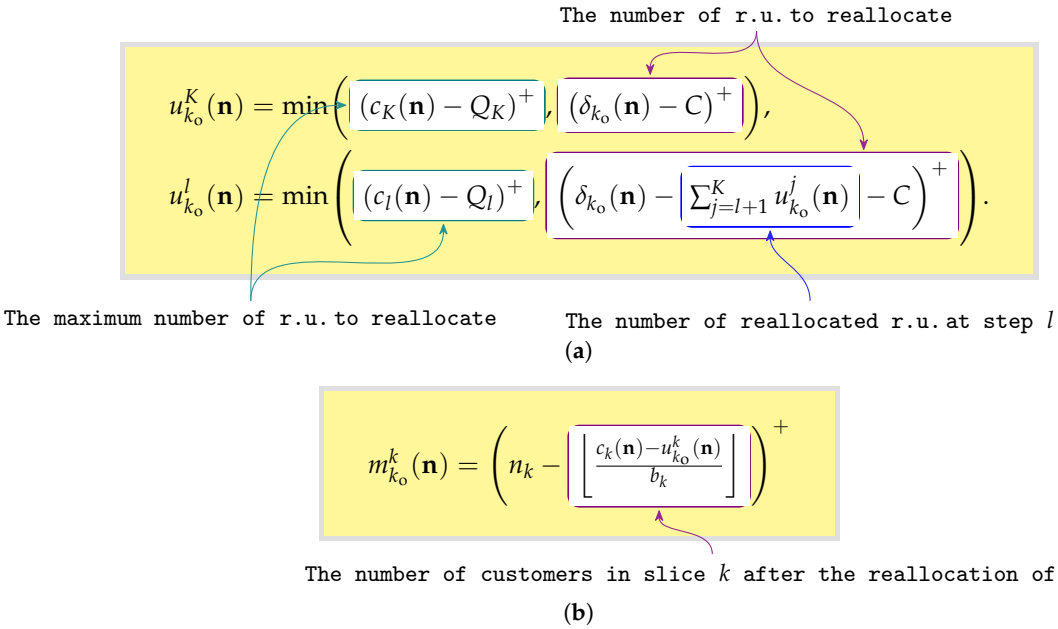


Figure 4. Explanations for Definition 2 (Dynamic Resource Reallocation (DRR)). (**a**) Details about Equations (10) and (11); (**b**) Details about Equation (12).

Given the above-described RAC scheme and the analytical DRS and DRR methods as fundaments of the second version of the Preemption-based service Prioritization (PP2), we can move to the mathematical model construction in the next section.

3. Mathematical Model

According to the Radio Admission Control (RAC) scheme given in Section 2 and considering the Poisson flows of customers, the system behavior can be described by the K-dimensional Markov process $\mathbf{X}(t) = \left(X_1(t), \ldots, X_K(t), t > 0\right)^T$, where $X_k(t), k \in \mathcal{K}$, is the number of customers in slice k at time t over the system state space

$$\mathcal{S} = \left\{ \mathbf{n} \in \Omega : \bigwedge_{k_e \in \mathcal{K}_e^+(\mathbf{n})} \left(x_{k_e}(\mathbf{n}) \geq b_{k_e}\right), \sum_{k=1}^{K} (n_k - M_k)^+ = 0, \sum_{k=1}^{K} c_k(\mathbf{n}) \leq C \right\}. \quad (13)$$

We note that system state space $\mathcal{S}$ contains several subspaces ([19], pp. 66–100), ([20], pp. 2–14). For a slice $k \in \mathcal{K}$, we denote the following:
$\mathcal{S}_k^{(1)}$—the state subspace for a direct admission (no preemption) of a customer;
$\mathcal{S}_k^{(2)}$—the state subspace for an admission via preemption;
$\overline{\mathcal{S}}_k$—the state subspace for blocking (or rejection):

$$\mathcal{S}_k^{(1)} = \left\{ \mathbf{n} \in \mathcal{S} : \bigwedge_{k_e \in \mathcal{K}_e^+(\mathbf{n}+\mathbf{e}_k)} \left(x_{k_e}(\mathbf{n}+\mathbf{e}_k) \geq b_{k_e}\right), n_k < M_k \right\}, \quad (14)$$

$$\mathcal{S}_k^{(2)} = \left\{ \mathbf{n} \in \mathcal{S} : \bigvee_{k_e \in \mathcal{K}_e^+(\mathbf{n}+\mathbf{e}_k)} \left(x_{k_e}(\mathbf{n}+\mathbf{e}_k) < b_{k_e}\right), n_k < N_k \right\}, \quad (15)$$

$$\overline{\mathcal{S}}_k = \left\{ \mathbf{n} \in \mathcal{S} : \bigvee_{k_e \in \mathcal{K}_e^+(\mathbf{n}+\mathbf{e}_k)} \left(x_{k_e}(\mathbf{n}+\mathbf{e}_k) < b_{k_e}\right), n_k + 1 > N_k \vee n_k = M_k \right\}, \quad (16)$$

$$\mathcal{S} = \mathcal{S}_k^{(1)} \cup \mathcal{S}_k^{(2)} \cup \overline{\mathcal{S}}_k. \quad (17)$$

Given the above-defined state space $\mathcal{S}$ and its subspaces, the RAC function for the customers flow $k \in \mathcal{K}$ can be described as $f_k(\mathbf{n}) = \begin{cases} 1, & \mathbf{n} \in \mathcal{S}_k^{(1)} \cup \mathcal{S}_k^{(2)}, \\ 0, & \text{otherwise } (\mathbf{n} \in \overline{\mathcal{S}}_k), \end{cases}, \mathbf{n} \in \mathcal{S}$.

Let $\mathbf{m}_k(\mathbf{n}) := \left(m_k^1(\mathbf{n}), \ldots, m_k^K(\mathbf{n})\right)^T, k \in \mathcal{K}, \mathbf{n} \in \mathcal{S}$.

A depiction of the system state transition diagram between $\mathbf{n} \in \mathcal{S}$ and all its neighbors $\mathbf{n} + \mathbf{e}_k, \mathbf{n} - \mathbf{e}_k, \mathbf{n} + \mathbf{e}_k - \mathbf{m}_k(\mathbf{n}), \mathbf{n} - \mathbf{e}_k + \mathbf{m}_k(\mathbf{n})$ is provided in Figure 5.

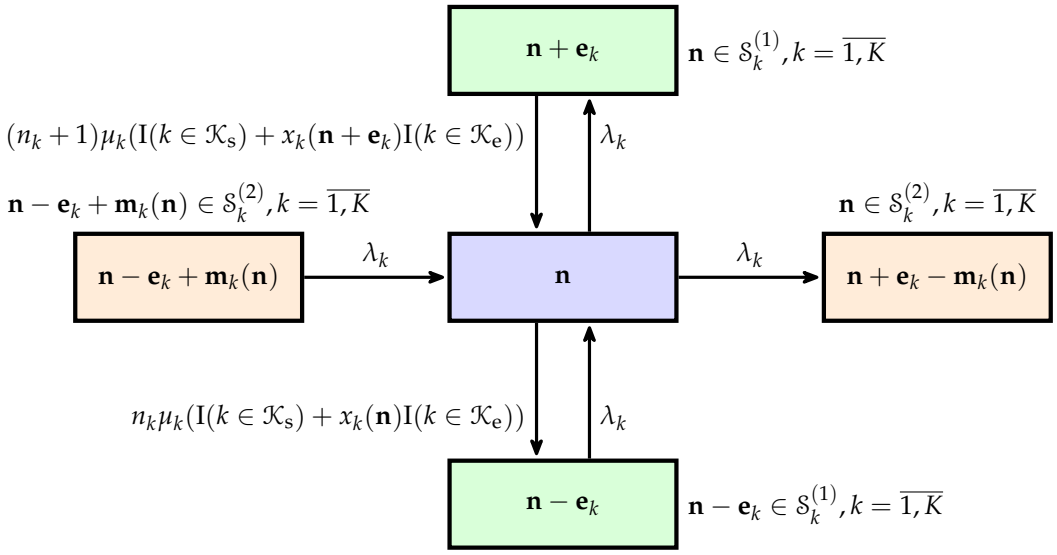


Figure 5. State transition diagram between $\mathbf{n} \in \mathcal{S}$ and its neighbors $\mathbf{n} \pm \mathbf{e}_k, \mathbf{n} \pm \mathbf{e}_k \pm \mathbf{m}_k(\mathbf{n})$.

Let $\mathbf{p}^T := \left(p(\mathbf{n})\right)_{\mathbf{n}\in\mathcal{S}}$ denote the stationary probability distribution of the Markov process $\mathbf{X}(t)$, with $p(\mathbf{n}), \mathbf{n} \in \mathcal{S}$, denoting the stationary probability that the system is in state $\mathbf{n}$. We can describe $\mathbf{p}^T$ using a system of equilibrium equations:

$$p(\mathbf{n})\left(\sum_{k=1}^{K} n_k\mu_k(\mathrm{I}(k\in\mathcal{K}_\mathrm{s}) + x_k(\mathbf{n})\mathrm{I}(k\in\mathcal{K}_\mathrm{e})) + \sum_{k=1}^{K}\lambda_k\mathrm{I}\left(\mathbf{n}\in\mathcal{S}_k^{(1)}\right) + \sum_{k=1}^{K}\lambda_k\mathrm{I}\left(\mathbf{n}\in\mathcal{S}_k^{(2)}\right)\right) =$$
$$= \sum_{k=1}^{K} p(\mathbf{n}-\mathbf{e}_k)\lambda_k\mathrm{I}\left(\mathbf{n}-\mathbf{e}_k\in\mathcal{S}_k^{(1)}\right) + \sum_{k=1}^{K} p(\mathbf{n}-\mathbf{e}_k+\mathbf{m}_k(\mathbf{n}))\lambda_k\mathrm{I}\left(\mathbf{n}-\mathbf{e}_k+\mathbf{m}_k(\mathbf{n})\in\mathcal{S}_k^{(2)}\right)+$$
$$+\sum_{k=1}^{K} p(\mathbf{n}+\mathbf{e}_k)(n_k+1)\mu_k(\mathrm{I}(k\in\mathcal{K}_\mathrm{s}) + x_k(\mathbf{n}+\mathbf{e}_k)\mathrm{I}(k\in\mathcal{K}_\mathrm{e}))\mathrm{I}\left(\mathbf{n}\in\mathcal{S}_k^{(1)}\right), \quad (18)$$

with the intrinsic normalization condition $\sum_{\mathbf{n}\in\mathcal{S}} p(\mathbf{n}) = 1$.

Due to the implementation of the Dynamic Resource Reallocation (DRR) method with service preemption, the Markov process $\mathbf{X}(t)$ is not reversible. Therefore, we can numerically compute the stationary probability distribution $\mathbf{p}^T$ using an iterative method ([21], pp. 103–204) ([22], pp. 106–179) [23–25] to solve Equation (18) rewritten as $\mathbf{p}^T\mathbf{A} = \mathbf{0}^T$, $\mathbf{p}^T\mathbf{1} = 1$, where $\mathbf{A} = \left(a(\mathbf{n},\mathbf{s})\right)_{\mathbf{n},\mathbf{s}\in\mathcal{S}}$ is the infinitesimal matrix (the matrix of transition intensities):

$$a(\mathbf{n},\mathbf{s}) = \begin{cases} \lambda_k,\ \mathbf{s} = \mathbf{n}+\mathbf{e}_k,\ \mathbf{n}\in\mathcal{S}_k^{(1)},\ k=\overline{1,K}, \\ \lambda_k,\ \mathbf{s} = \mathbf{n}+\mathbf{e}_k-\mathbf{m}_k(\mathbf{n}),\ \mathbf{n}\in\mathcal{S}_k^{(2)},\ k=\overline{1,K}, \\ n_k\mu_k(\mathrm{I}(k\in\mathcal{K}_\mathrm{s}) + x_k(\mathbf{n})\mathrm{I}(k\in\mathcal{K}_\mathrm{e})),\ \mathbf{s} = \mathbf{n}-\mathbf{e}_k, k=\overline{1,K}, \\ -\sum_{k=1}^{K}\lambda_k\mathrm{I}\left(\mathbf{n}\in\mathcal{S}_k^{(1)}\right) - \sum_{k=1}^{K}\lambda_k\mathrm{I}\left(\mathbf{n}\in\mathcal{S}_k^{(2)}\right) - \sum_{k=1}^{K} n_k\mu_k(\mathrm{I}(k\in\mathcal{K}_\mathrm{s}) + x_k(\mathbf{n})\mathrm{I}(k\in\mathcal{K}_\mathrm{e})),\ \mathbf{s} = \mathbf{n}, \\ 0,\ \text{otherwise.} \end{cases} \quad (19)$$

Once the stationary probability distribution $\mathbf{p}^T$ is obtained, one can compute the system Key Performance Indicators (KPIs) using expressions in analytic form. We analyze the system functioning in terms of the utilization and accessibility KPIs [3]. Let us denote the following:

$N_k, k\in\mathcal{K}$—the mean number of customers in slice k;
N—the mean number of customers in the system;
P_k—the admission probability at slice k (the probability of an event $\mathcal{S}_k^{(1)}\cup\mathcal{S}_k^{(2)}$);
P—the admission probability at the system;
ϕ_{Q_k}—the guaranteed capacity utilization in slice k;
$\phi_{b_{k_\mathrm{e}}}, k_\mathrm{e}\in\mathcal{K}_\mathrm{e}$—the minimum data rate utilization in TCP elastic slice k_e;
ϕ_C—the total system capacity utilization:

$$N_k = \sum_{\mathbf{n}\in\mathcal{S}} p(\mathbf{n})n_k, \quad k\in\mathcal{K}, \quad (20)$$

$$N = \sum_{\mathbf{n}\in\mathcal{S}} p(\mathbf{n})\sum_{k=1}^{K} n_k = \sum_{k=1}^{K} N_k, \quad (21)$$

$$\mathrm{P}_k = \sum_{\mathbf{n}\in\mathcal{S}_k^{(1)}\cup\mathcal{S}_k^{(2)}} p(\mathbf{n}) = \sum_{\mathbf{n}\in\mathcal{S}\setminus\overline{\mathcal{S}}_k} p(\mathbf{n}), \quad (22)$$

$$\mathrm{P} = \sum_{\mathbf{n}\in\mathcal{S}\setminus\bigcap_{k=1}^{K}\overline{\mathcal{S}}_k} p(\mathbf{n}), \quad (23)$$

$$\phi_{Q_k} = \frac{1}{Q_k}\sum_{\mathbf{n}\in\mathcal{S}} p(\mathbf{n})c_k(\mathbf{n}), \quad (24)$$

$$\phi_{b_{k_e}} = \frac{1}{b_{k_e}} \sum_{\mathbf{n}\in\mathcal{S}} p(\mathbf{n}) x_{k_e}(\mathbf{n}), \quad k_e \in \mathcal{K}_e, \tag{25}$$

$$\phi_C = \frac{1}{C} \sum_{\mathbf{n}\in\mathcal{S}} p(\mathbf{n}) \sum_{k=1}^{K} c_k(\mathbf{n}). \tag{26}$$

Once the KPIs are computed, one can move to their analysis in the next section.

4. Numerical Analysis

In this section, we provide a comparative analysis of the model developed in Sections 2 and 3, which is the second version of the Preemption-based service Prioritization (PP2) scheme, versus the Resource Reservation (RR) [10]. We use the expression $\Delta := \left|\frac{\text{KPI}_{\text{PP2}} - \text{KPI}_{\text{RR}}}{\text{KPI}_{\text{RR}}}\right| \times 100\%$ to denote the relative gain or loss in terms of a given Key Performance Indicator (KPI).

Let us consider a case example of three slices coexisting in a system ($K = 3$). Let $C = 100$ resource units (r.u.). For simplicity, we assume $C_k, Q_k, b_k, k = 1, 2, 3$, are expressed in percentages of number C [26,27]. Let us regroup the input parameters in Table 2.

Table 2. Input data for KPI analysis, given the uniform offered loads $\rho_k = \rho = \overline{1,300}, k = 1, 2, 3$.

Slice (Traffic)	Parameter	Value RR Scheme	Value PP2 Scheme	Unit
1, 2 (Streaming)	$C_1, Q_1, b_1, C_2, Q_2, b_2$	$35, 35, 4, 35, 35, 3$	$40, 35, 4, 40, 35, 3$	% of C
	$\mu_1, \mu_2, \lambda_1, \lambda_2$	$60, 30, \rho\mu_1, \rho\mu_2$		customers/time units (t.u.)
3 (Elastic)	C_3, Q_3, b_3	$30, 30, 5$	$75, 30, 5$	% of C
	μ_3^{-1}	10		size units (s.u.)
	λ_3	$\rho\mu_1$		customers/t.u.

As one can guess from Table 2, we compared the classical RR scheme (Complete Partitioning, where $C_k = Q_k, k = 1, 2, 3$, with $C_1 + C_2 + C_3 = 100\%$ of C) [5,6,8] and the PP2 scheme with an overlap of 5% of C for each non-TCP streaming slice 1, 2 and 45% of C for TCP elastic slice 3. The Priority Levels (PLs) were set as shown in the first column of Table 1 to make non-TCP streaming slices 1, 2 gain the most from the Dynamic Resource Reallocation (DRR) method with data rate reduction (Definition 2). Note that the PLs are used only under PP2. The comparison results for the utilization and accessibility KPIs described in Equations (20)–(26) are depicted in Figures 6–8.

The main results of the comparative analysis are summarized as follows. The PP2 scheme application under the baseline parameters in Table 2 shows several gains (✓) and one loss (✗) in terms of utilization and accessibility KPIs:

✓ The mean number of customers in non-TCP streaming slice 1 can increase up to 28% (Figure 6a);
✓ The mean number of customers in non-TCP streaming slice 2 can increase up to 19.75% (Figure 6b);
✓ The mean number of customers in TCP elastic slice 3 can decrease down to 40% (Figure 6c);
✓ The mean number of customers in system can increase up to 9% (Figure 6d);
✓ The admission probability of customers at non-TCP streaming slice 1 can increase up to 17.5% (Figure 7a);
✓ The admission probability of customers at non-TCP streaming slice 2 can increase up to 15% (Figure 7b);
✓ The admission probability of customers at TCP elastic slice 3 can increase up to 99.999% (Figure 7c);
✓ The admission probability of customers at the system can increase up to 99.999% (Figure 7d);

✓ The guaranteed capacity utilization in non-TCP streaming slice 1 can increase up to 28% (Figure 8a);
✓ The guaranteed capacity utilization in non-TCP streaming slice 2 can increase up to 19.75% (Figure 8b);
✗ The guaranteed capacity utilization in TCP elastic slice 3 can decrease down to 30% (Figure 8c);
✓ The minimum data rate utilization in TCP elastic slice 3 can increase up to 52% (Figure 8d);
✓ The total system capacity utilization can increase up to 13% (Figure 8e).

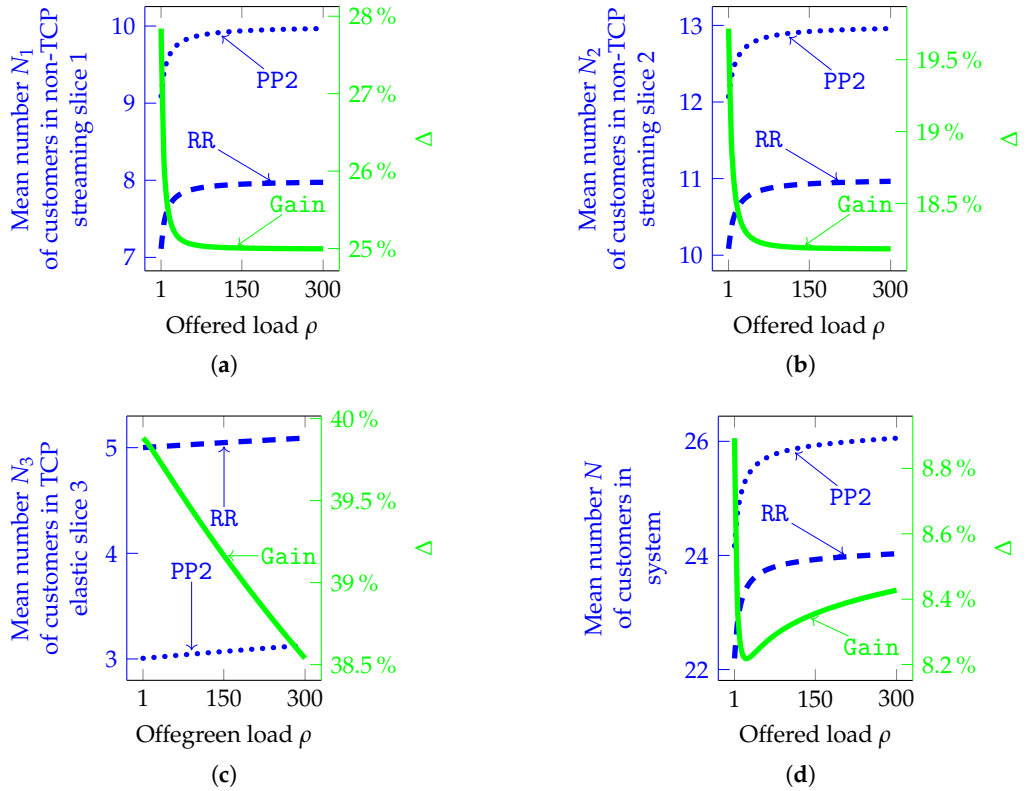


Figure 6. The numerical results of a case example of three slices in a system ($K = 3$) for the mean number of customers vs. the offered load. (**a**) The mean number of customers in non-TCP streaming slice 1; (**b**) The mean number of customers in non-TCP streaming slice 2; (**c**) The mean number of customers in TCP elastic slice 3; (**d**) The mean number of customers in the system.

The obtained results are analyzed as follows. Under the PP2 scheme, non-TCP streaming slices 1, 2 can service more customers than guaranteed—shown by the PP2 utilization curve above the horizontal unit in Figure 8a,b, which is corroborated by the PP2 curve above the RR curve in Figure 6a,b. This is due to the DRR method reducing (down toward the minimum data rate) the data rate of each customer in TCP elastic slice 3—shown by the PP2 utilization curve below the horizontal unit in Figure 8c. Even with this, as expected by allowing an overlap of 45% of C for TCP elastic slice 3 (Table 2), each of its customers can be allocated almost the double of the minimum data rate—shown by the PP2 utilization curve above the horizontal unit and above the RR utilization curve in Figure 8d, which leads to fast service—shown by the PP2 curve below the RR curve in Figure 6c. At last, the system capacity is constantly 100% utilized—shown by the PP2 utilization curve remaining at the horizontal unit in Figure 8e.

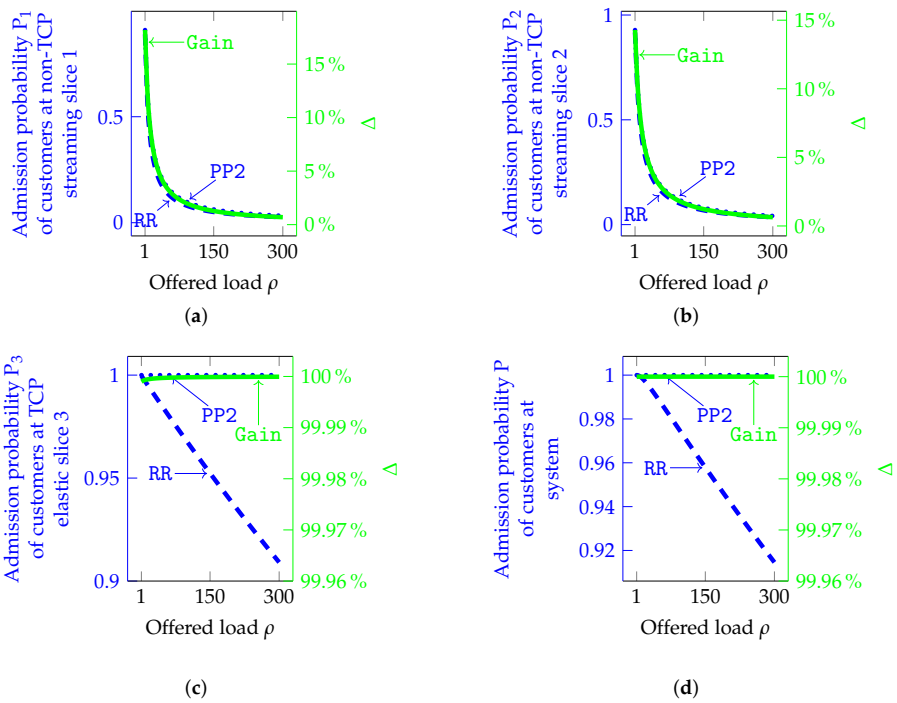


Figure 7. The numerical results of a case example of three slices in a system ($K = 3$) for the admission probability of customers vs. the offered load. (**a**) The admission probability of customers at non-TCP streaming slice 1; (**b**) The admission probability of customers at non-TCP streaming slice 2; (**c**) The admission probability of customers at TCP elastic slice 3; (**d**) The admission probability of customers at the system.

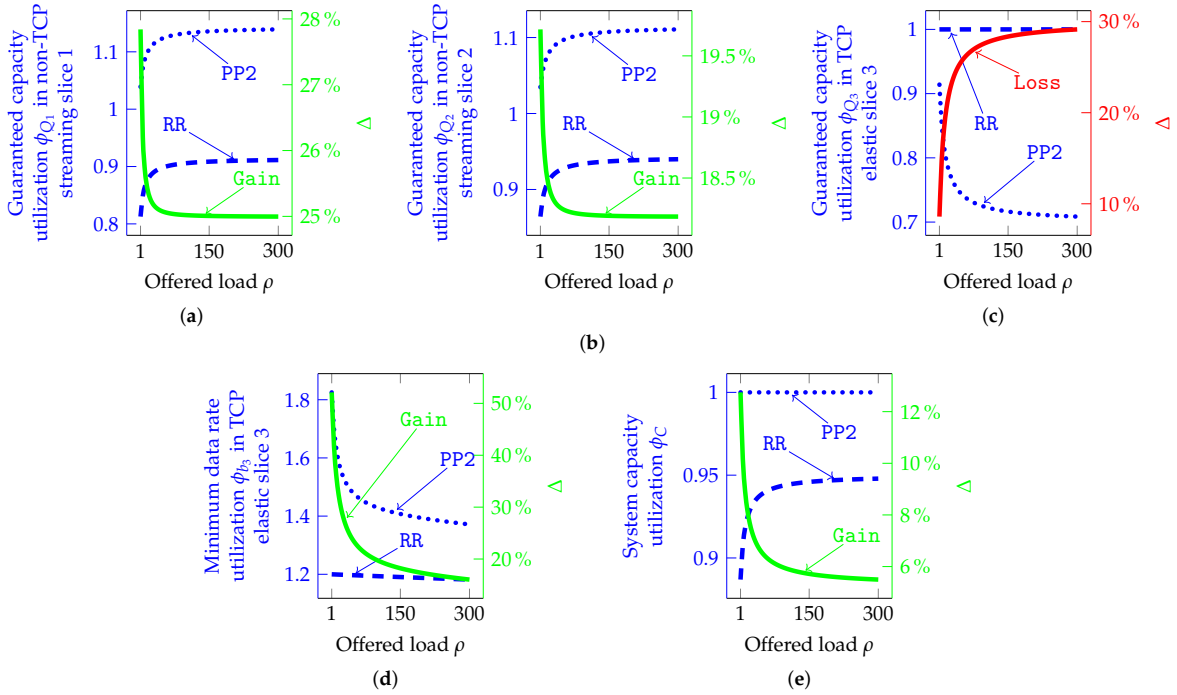


Figure 8. The numerical results of a case example of three slices in a system ($K = 3$) for the capacity utilization vs. the offered load. (**a**) The guaranteed capacity utilization in non-TCP streaming slice 1; (**b**) The guaranteed capacity utilization in non-TCP streaming slice 2; (**c**) The guaranteed capacity utilization in TCP elastic slice 3; (**d**) The minimum data rate utilization in TCP elastic slice 3; (**e**) The total system capacity utilization.

5. Conclusions

The current research developed the second version of the Preemption-based service Prioritization (PP2) scheme—an improvement of the previous PP [10]. Unlike PP, which optimizes resource utilization among slices, with each providing non-TCP streaming service (non-TCP streaming slice), PP2 realizes the same among slices, each providing either non-TCP streaming or TCP elastic service (TCP elastic slice). PP2 is based on analytically described Dynamic Resource Sharing (DRS) and Dynamic Resource Reallocation (DRR) methods—DRS ensures fair resource allocation, while DRR reduces the capacity share of the slices violating the mandatory isolation among the slices (slice violators). The main feature of PP2 is the mutual utilization of the guaranteed capacity of all slices. Given some baseline scenario, the comparison of PP2 versus the classical Resource Reservation (RR) shows (i) an increase up to 99.999% in the admission probability at the system and (ii) a constant system capacity utilization at 100%.

As a next research topic, one can consider the application of Service Prioritization (SP) schemes and Priority Levels (PLs) on all TCP elastic slices in system capacity utilization. As another interesting subject, one can research the benefits of introducing the maximum data rate of customers with less than the overall capacity into all TCP elastic slices.

In general, unlike the previous PP scheme, PP2 can be used at various levels in smart cities, farms, and buildings to efficiently manage or supply available water, energy, etc.

Author Contributions: Conceptualization, Y.A., E.M. and Y.G.; methodology, Y.A., E.M. and Y.G.; software, Y.A.; validation, Y.A.; formal analysis, Y.A.; investigation, Y.A.; resources, Y.A.; data curation, Y.A.; writing—original draft preparation, Y.A.; writing—review and editing, Y.A., E.M. and Y.G.; visualization, Y.A. and E.M.; supervision, Y.A., E.M. and Y.G.; project administration, Y.A., E.M. and Y.G.; funding acquisition, E.M. and Y.G. 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. 22-79-10053 (https://rscf.ru/en/project/22-79-10053/, accessed on 24 August 2024) (recipient Markova, Sections 2 and 3). This publication has been supported by the RUDN University Scientific Projects Grant System, project no. 021937-2-000 (recipient Gaidamaka, Sections 1 and 4).

Data Availability Statement: The generated research data are publicly available from FigShare at https://doi.org/10.6084/m9.figshare.c.7404859.v1 (accessed on 10 October 2024).

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

Abbreviations

The following abbreviations are used in this manuscript:

3GPP	Third Generation Partnership Project
DRR	Dynamic Resource Reallocation
DRS	Dynamic Resource Sharing
gNB	Fifth Generation Base Station
KPI	Key Performance Indicator
NS	Network Slicing
PL	Priority Level
PP	Preemption-based service Prioritization
PP2	second version of the PP
QoS	Quality of Service
r.u.	resource units
RAC	Radio Admission Control
RR	Resource Reservation
s.u.	size units
SP	Service Prioritization
t.u.	time units

References

1. Chandramouli, D.; Pope, M. 5G; System architecture for the 5G System (5GS) (3GPP TS 23.501 version 17.9.0 Release 17). Available online: https://www.etsi.org/deliver/etsi_ts/123500_123599/123501/17.09.00_60/ts_123501v170900p.pdf (accessed on 10 October 2024).
2. Sultan, A.; Pope, M. Digital cellular telecommunications system (Phase 2+) (GSM); Universal Mobile Telecommunications System (UMTS); LTE; 5G; Release 16 Description; Summary of Rel-16 Work Items (3GPP TR 21.916 version 16.2.0 Release 16). Available online: https://www.etsi.org/deliver/etsi_tr/121900_121999/121916/16.00.01_60/tr_121916v160001p.pdf (accessed on 10 October 2024).
3. Sun, X.; Soveri, M.C. 5G; Management and orchestration; 5G end to end Key Performance Indicators (KPI) (3GPP TS 28.554 version 17.10.0 Release 17). Available online: https://www.etsi.org/deliver/etsi_ts/128500_128599/128554/17.10.00_60/ts_128554v171000p.pdf (accessed on 10 October 2024).
4. Li, A.; Sultan, A. 5G; Service requirements for the 5G system (3GPP TS 22.261 version 17.11.0 Release 17). Available online: https://www.etsi.org/deliver/etsi_ts/122200_122299/122261/17.11.00_60/ts_122261v171100p.pdf (accessed on 10 October 2024).
5. Vila, I.; Sallent, O.; Umbert, A.; Perez-Romero, J. An Analytical Model for Multi-Tenant Radio Access Networks Supporting Guaranteed Bit Rate Services. *IEEE Access* **2019**, *7*, 57651–57662. [CrossRef]
6. Banchs, A.; de Veciana, G.; Sciancalepore, V.; Costa-Perez, X. Resource Allocation for Network Slicing in Mobile Networks. *IEEE Access* **2020**, *8*, 214696–214706. [CrossRef]
7. Narmanlioglu, O.; Zeydan, E.; Arslan, S.S. Service-Aware Multi-Resource Allocation in Software-Defined Next Generation Cellular Networks. *IEEE Access* **2018**, *6*, 20348–20363. [CrossRef]
8. Fu, X.; Shen, Q.; Wang, W.; Hou, H.; Gao, X. Slice Merging/Spliting Operations and Tenant Profit Optimization Across 5G Base Stations. *IEEE Access* **2021**, *9*, 9706–9718. [CrossRef]
9. Zhou, X.; Kim, D. 5G; 5G System; Session Management Policy Control Service; Stage 3 (3GPP TS 29.512 version 17.11.0 Release 17). Available online: https://cdn.standards.iteh.ai/samples/68674/5f5cb7045e6c4dbcac881e022b0a43d2/ETSI-TS-129-512-V17-11-0-2023-07-.pdf (accessed on 10 October 2024).
10. Adou, Y.; Markova, E.; Gaidamaka, Y. Modeling and Analyzing Preemption-Based Service Prioritization in 5G Networks Slicing Framework. *Future Internet* **2022**, *14*, 299. [CrossRef]
11. Yarkina, N.; Gaidamaka, Y.; Correia, L.M.; Samouylov, K. An analytical model for 5G network resource sharing with flexible SLA-oriented slice isolation. *Mathematics* **2020**, *8*, 1177. [CrossRef]
12. Gonçalves, D.M.; Bittencourt, L.F.; Madeira, E.R. Overhead and Performance of Dynamic Network Slice Allocation for Mobile Users. *Future Gener. Comput. Syst.* **2024**, *160*, 739–751. [CrossRef]
13. Camargo, J.S.; Coronado, E.; Ramirez, W.; Camps, D.; Deutsch, S.S.; Pérez-Romero, J.; Antonopoulos, A.; Trullols-Cruces, O.; Gonzalez-Diaz, S.; Otura, B.; et al. Dynamic slicing reconfiguration for virtualized 5G networks using ML forecasting of computing capacity. *Comput. Netw.* **2023**, *236*, 110001. [CrossRef]
14. Khodapanah, B.; Awada, A.; Viering, I.; Francis, J.; Simsek, M.; Fettweis, G.P. Radio Resource Management in context of Network Slicing: What is Missing in Existing Mechanisms? In Proceedings of the 2019 IEEE Wireless Communications and Networking Conference (WCNC), Marrakesh, Morocco, 15–18 April 2019; pp. 1–7. [CrossRef]
15. 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]
16. Key, P.; Massoulié, L.; Bain, A.; Kelly, F. A network flow model for mixtures of file transfers and streaming traffic. In *Teletraffic Science and Engineering*; Elsevier: Amsterdam, The Netherlands, 2003; pp. 1021–1030. [CrossRef]
17. Kelly, F. Charging and rate control for elastic traffic. *Eur. Trans. Telecommun.* **1997**, *8*, 33–37. [CrossRef]
18. Bertsekas, D.P.; Gallager, R.G. *Data Networks*, 2nd ed.; Pearson: Upper Saddle River, NJ, USA, 1992; p. 576.
19. Devlin, K. *The Joy of Sets*; Springer: New York, NY, USA, 1993; p. 194. [CrossRef]
20. Bagaria, J. Set Theory. Available online: https://plato.stanford.edu/archives/spr2023/entries/set-theory/ (accessed on 24 August 2024).
21. Bocharov, P.P.; D'Apice, C.; Pechinkin, A.V. *Queueing Theory*; De Gruyter: Berlin, Germany, 2003; p. 446. [CrossRef]
22. Stepanov, S.N. *Teoriya Teletrafika: Kontseptsii, Modeli, Prilozheniya (Teletraffic Theory: Concepts, Models, Applications)*; Goryachaya Liniya—Telekom: Moscow, Russia, 2015; p. 868.
23. Basharin, G.P.; Gaidamaka, Y.V.; Samouylov, K.E. Mathematical theory of teletraffic and its application to the analysis of multiservice communication of next generation networks. *Autom. Control Comput. Sci.* **2013**, *47*, 62–69. [CrossRef]
24. Basharin, G.P.; Tolmachev, A.L. Theory of queuing networks and its applications to the analysis of information-computing systems. *J. Math. Sci.* **1985**, *29*, 951–1050. [CrossRef]
25. Basharin, G.P.; Samouylov, K.E.; Yarkina, N.V.; Gudkova, I.A. A new stage in mathematical teletraffic theory. *Autom. Remote Control* **2009**, *70*, 1954–1964. [CrossRef]

26. Standaert, G.; Uitto, T. Nokia and Proximus Deliver 5G Network Slicing Innovation to Support Demanding Network Conditions. Available online: https://www.nokia.com/about-us/news/releases/2022/06/28/nokia-and-proximus-deliver-5g-network-slicing-innovation-to-support-demanding-network-conditions/ (accessed on 24 August 2024).
27. Kynäslahti, A. Nokia and Google Trial Innovative Slicing Solution for Android Devices and 4G/5G Networks. Available online: https://www.nokia.com/about-us/news/releases/2022/08/25/nokia-and-google-trial-innovative-slicing-solution-for-android-devices-and-4g5g-networks/ (accessed on 24 August 2024).

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.