

SLA-Driven Radio Resource Management Using Control Parameter Optimization in 5G Network Slicing

Qasim Abduljabbar Hamad¹, Morteza Valizadeh^{2, *}, Vahid Talavat³

q.abduljabbarhamad@urmia.ac.ir, mo.valizadeh@urmia.ac.ir, v.talavat@urmia.ac.ir

^{1, 2, 3} Department of electrical and computer engineering, Urmia University, Urmia, Iran.

*Corresponding author: Morteza Valizadeh, mo.valizadeh@urmia.ac.ir

ARTICLE INFO

Received: 18 Dec 2024

Revised: 14 Feb 2025

Accepted: 22 Feb 2025

ABSTRACT

In traditional telecommunications networks, introducing new functions or processes, especially with diverse approaches to quality assurance, often requires fundamental changes to the architecture and configuration of existing networks, which will lead to hardware modifications. Consequently, such changes usually faced widespread resistance from operators due to the significant costs involved. A viable solution that gained widespread popularity in the late first decade of this century is the software-centric redesign of network functions to facilitate centralized management, reduce expansion and update costs, and enhance the overall efficiency of the system. Alongside extensive academic research in this direction, this solution is now significantly implemented in the industry, particularly in fifth-generation communications and beyond. The diverse and potentially conflicting requirements of new applications in modern wireless telecommunications present a significant challenge in maintaining connectivity and coherence among various blocks of an integrated network to meet the quality needs of diverse applications. One proposed solution in this field, based on the approach of software-defined networks, is network slicing. The main concept of slicing involves sharing various physical network resources over virtual networks with centralized management and predefined quality of service requirements. For this purpose, different network slices must be managed and configured by a central control unit. In the cellular wireless network studied in this research, this unit is introduced as the mapping layer. This layer monitors its serviced network and manages the allocation of radio resources to slices based on a specific method to meet the service needs of each slice. Following an initial introduction, the proposed idea is compared with some existing and new methods. Simulation results from this research indicate that the proposed slicing method performs better in terms of meeting key performance indicators compared to other methods reviewed, especially when the demand for resources exceeds the capacity allocated to a slice, relative to the agreed service level.

Keywords: *Key performance indicators, Service level agreement (SLA), Fifth-generation networks, Resource management, Mapping layer, Cellular network, Slicing.*

1- INTRODUCTION

The advent of 5G networks brings forward the promise of delivering diverse and demanding services, from high-speed mobile broadband to ultra-reliable low-latency communications and massive machine-type communications. A pivotal technology enabling these services is network slicing, which allows the partitioning of a single physical network into multiple virtual networks, each tailored to meet the specific needs of a service or application. Deep learning, with its ability to learn complex patterns

and make intelligent decisions, emerges as a key player in optimizing resource allocation and managing network slices efficiently. The benefits, trends, and challenges associated with the softwarization and virtualization of 5G mobile networks are comprehensively discussed, highlighting the foundation for network slicing [1]. The role of slice-aware radio resource management in fulfilling service level agreements is further elaborated, illustrating the practical applications and operational benefits of network slicing [2]. Standardization efforts by industry consortia and organizations, such as GSMA [3], NGMN Alliance [4], and 3GPP [5], underscore the global momentum towards adopting network slicing as a core 5G technology, outlining use case requirements, concepts, and technical specifications that underpin network slicing architectures. The necessity for refined radio resource management mechanisms in the context of network slicing is analyzed in [6], pointing to gaps in existing frameworks and proposing directions for future research. Empirical studies, such as, assess the impact of network slicing on 5G radio access networks [7], providing valuable insights into its operational implications. The final overall 5G RAN design, as delivered by 3GPP [8], sets a comprehensive blueprint for implementing network slicing within the radio access network, guiding the industry towards standardized deployment models. Resource allocation within the context of network slicing presents unique challenges, particularly in ensuring efficient utilization of network resources while catering to the varied requirements of different slices. The optimization models and algorithms proposed in [9] address joint uplink/downlink planning, highlighting the criticality of effective resource management strategies. Surveys and studies [10], [11] explore the broader implications of resource slicing and wireless network virtualization, offering a theoretical foundation and identifying research issues and challenges that persist in the field. The advent of deep learning in network slicing and resource allocation offers novel pathways to address these challenges. Deep reinforcement learning, in particular, has been identified as a potent tool for dynamic and autonomous network management [16], [17], [28], [29]. These studies demonstrate the application of deep learning techniques in optimizing slice management and radio resource allocation, showcasing the potential for significant improvements in network performance and efficiency. The iterative adaptation strategies for slice management [18], along with auction-based models for network slicing [20], exemplify innovative approaches to resource allocation that leverage the analytical capabilities of deep learning. Understanding the requirements of network slicing is essential for its implementation and optimization. The GSMA and NGMN have provided insights into use case requirements and the conceptual framework for network slicing, highlighting its significance for future mobile networks [3,4]. The management and orchestration of network slices are further detailed in 3GPP's technical specifications, which outline the concepts, use cases, and requirements for slice management [5]. However, fulfilling the promises of network slicing requires addressing several challenges, particularly in radio resource management (RRM). Studies have identified gaps in existing RRM mechanisms concerning network slicing and proposed enhancements to ensure efficient utilization of resources across slices [6,7]. Deep reinforcement learning (DRL) has emerged as a promising approach to tackle these challenges by enabling adaptive and intelligent slice management [16,28,29]. The application of DRL in network slicing encompasses various aspects, from slice-aware radio resource management to dynamic slice adjustment based on real-time network conditions [16,17,18]. By learning from the network's operational data, DRL models can predict and react to changes, ensuring optimal resource allocation and SLA fulfillment. Moreover, research has also explored auction-based models for network slicing, which introduce market-driven mechanisms for resource allocation, further enhancing the efficiency and fairness of slice management [19,20]. Service-aware multi-resource allocation strategies have been proposed to ensure that the diverse requirements of next-generation cellular networks are met, emphasizing the need for intelligent and flexible resource management solutions [21].

Due to the increasing proliferation of applications and the need for QoS (Quality of Service) and QoE (Quality of Experience) in fifth-generation cellular networks, the need for integrated and intelligent management of these networks has doubled. Fifth-generation and higher cellular networks should be able to support multiple service classes with diverse requirements such as enhanced mobile broadband

(eMBB), massive machine-type communications (mMTC), and ultra-reliable low latency communications (URLLC) [1]. It is essential to prevent the establishment of various networks separately or in parallel for each of the mentioned cases, as this would lead to inappropriate use of network resources and inflexible resource allocation, which would be counterproductive in network software-centric management and configuration. Therefore, a cost-effective, flexible, and scalable solution is necessary that can be used for the different services mentioned. A software-centric approach, such as slicing or using slicing, is a viable solution [2]. Initially, a comprehensive management framework is introduced with the presentation of the mapping layer and network controller. Two Key Performance Indicators (KPIs) are suggested for evaluating the performance of the network controller [2]. By thoroughly examining the performance of the mapping layer and the concept of slicing, and the necessity of isolating different slices in the service environment, various slicing methods were implemented by introducing three different models with various requirements [6]. Additionally, in the mentioned article, a new unit called the acceptance control was added to the mapping layer and its effect on different anomalies was investigated. The relationship between control parameters and changes in KPIs was investigated using a reinforcement learning framework and compared with various combined methods [16]. To achieve SLA compliance, a repetitive algorithm for improving control parameters AC and PS alongside the presence of anomalies such as traffic increases or congestion factors were introduced [18]. The compatibility of the proposed layer with the knowledge of the relationship between control parameters and KPIs, presented respectively using the reactive response matrix - Jacobian matrix - and ANN, needs to be examined. Investigating the impact of these relationships and understanding them has been studied; furthermore, a protective mechanism has been introduced that overlooks deviations in excess traffic and prioritizes sections with normal traffic loads [17].

Table 2: A summary comparison of the methods presented in this research.

Proposed method	Slice management in radio access network through iterative matching algorithm [17]	Management of radio resources in the form of slicing the network [6]	Realization of service level agreement through the management of slicing radio resources in fifth generation networks [2]	Research
Yes	Yes	Yes	Yes	The existence of an independent department that is in charge of controlling resources between slices
6	6	6	2	Number of problem of KPIs
3	3	3	2-4	Number of slices
Yes	Yes	Yes	No	Fair distribution of bandwidth
Yes - resources are allocated	Yes - congested slices	No	No	Prioritizing slices

according to the definition of SLA	have the lowest priority			
Pseudo-Jacobin matrix	Pseudo-Jacobin matrix and fit matrix based on ANN network	Pseudo-Jacobin matrix	A direct relationship based on the formula (2-8 and 2-9)	Relationship between control parameters and KPIs
Yes	Yes	Yes	No	Admission control
No	No	No	No	Mobility
7	21	7	7	The number of cells studied
Variable from uniform to Gaussian with the help of density factor	Variable from uniform to Gaussian with the help of density factor	Variable from uniform to Gaussian with the help of density factor	Gaussian or uniform	Spatial distribution of users
Local deviation matrix	Hadamard product of local deviation matrix in general	Hadamard product of local deviation matrix in general	Local deviation matrix	Calculation of KPI deviation from SLA
Variable, depending on network conditions	Constant	Constant	No	Update step factor

Highlights:

This study has presented a flexible and adjustable framework for RAN slicing where various requirements of slices are simultaneously considered, and slicing management algorithms adjust various parameters to control the management of radio resources to meet the service-level agreements of the slices. All of these algorithms and requirements have been thoroughly discussed in the previous section. With an overview of the adaptation methods for mapping layer discussed in previous studies, the aim here is to propose a method that introduces SLA profiles and prioritization to have better organization for updating control parameters in the presence of anomalies.

2- PROBLEM STATEMENT

In this study, radio resource management in a sliced network is investigated, where a new network entity called the KPI Mapping Layer monitors different slices and makes a series of changes based on the corresponding SLA. To begin with Packet scheduler which is part of the introduced entity, manages the allocation of each slice from the physical resource and ensures their respective SLA levels are met [8]. The Mapping Layer tracks the performance of slices through Key Performance Indicators (KPIs) in different cells and assigns suitable parameters to determine the slices' share of radio resources. Additionally, this entity must decide which slices in cells have a higher priority for resource allocation because in dense network conditions (where demand for resources is high), the mapping layer must be able to protect slices with low demand against the flood of demand for other slices (isolating slices from

each other). Here, an algorithm tailored to slice the mapping layer, attempting to minimize deviations from target KPIs in a service area against anomalies (increased incoming requests and high congestion) by allocating weights to slices. Understanding the following concepts is essential to comprehend this algorithm.

2-2-1. Proposed Model

Consider a cellular network with S sectors and C cells. The service area includes several neighboring cells that provide services to users through various cells. If the number of users in sector s in cell c is represented by $N_{s,c}$ and the total number of users in sector s in the service area is shown by: $N_{s,*} = \sum_{c=1}^C N_{s,c}$, the traffic load on sector s can be obtained through the following equation:

$$L_s = \frac{N_{s,*}}{a} \left[\frac{\text{user}}{\text{km}^2} \right] \quad (1)$$

A is the area which serves as a service region. If the traffic load in a section within the overall service area exceeds the predefined traffic load level in the Service Level Agreement (SLA), the network may not be able to achieve Key Performance Indicators (KPIs) for all sections, leading to network congestion. The network must prioritize other sections with lower traffic loads in this situation.

2-2-1. Package Manager

To prioritize different slices, a weight parameter is assigned to each slice and its users in each cell, and the package manager allocates its resources to the slices based on that weight. The weight matrix is defined as follows:

$$W = \begin{bmatrix} w_{1,1} & \dots & w_{1,C} \\ \vdots & \ddots & \vdots \\ w_{S,1} & \dots & w_{S,C} \end{bmatrix} = [W_{*,1} \quad W_{*,2} \quad \dots \quad W_{*,C}] \quad (2)$$

$W_{*,c}$ is a vector with dimensions $S \times 1$ that includes the weights of all S slices in cell C . Therefore, the contribution of user i to slices in cell C is calculated as follows:

$$r_{s,c}^i(W_{*,c}) = \frac{\omega_{s,c}}{\sum_{s'}^S N_{s',c} \cdot \omega_{s',c}} \quad (3)$$

Please note that in this equivalence, the weights are normalized. By having the contribution of each user, the throughput of each user can be calculated:

$$T_{s,c}^i(W_{*,c}) = r_{s,c}^i(W_{*,c}) \cdot n \cdot B \cdot \log_2(1 + \gamma_{s,c}^i) \quad (4)$$

$\gamma_{s,c}^i$ is the value of SINR (Signal to-Interference-Plus-Noise-Ratio) for user i determined by the number of physical resource blocks (PRBs), denoted as n , and B is the bandwidth of each PRBs. It is important to note that in this context, we employ Shannon's capacity formula to map SINR to the bit rate.

2-2-1-2. Definition of KPIs

There are a large number of KPIs that can be considered in an SLA. For example: delay, coverage level, energy efficiency, etc. Since this research focuses on studying the impact of PS on achieving the agreed service level, specific KPIs are introduced that can best demonstrate its impact in cellular networks. For this purpose, both average throughput and minimum throughput are considered as two KPIs. To calculate the average throughput of a slice in the network, averaging is performed across all cell users. For minimum throughput, first the minimum throughput in each cell is calculated and then on the entire cells:

$$As(W) = \left(\frac{\sum_{c=1}^C \sum_{i=1}^{\omega_{s,c}} T_{s,c}^i(W_{*,c})}{N_{s,*}} \right) \quad (5)$$

$$Ms(W) = \min c \in \{1.2. \dots c\} \quad (\min_i) \in \{1.2. \dots c\} T_{s,c}^i(W_{*,c}) \quad (6)$$

Each slice s aims at a mean and minimum permittivity goal, which is represented by \hat{A}_s , \hat{M}_s , and the final traffic limit of a slice L .

2-2-1-3. Optimization Algorithm for Mapping Layer

In this section, the aim is to utilize optimization algorithms to examine the impact of a central entity on achieving SLAs by considering $\Phi_s^A(X)$ and Φ_s^M , which will be defined in the next section, as cost functions related to the average and minimum throughput of slice S . The weighting of slices, as mentioned, lies with unit PS, which is a part of the mapping layer itself. The total cost function is defined as the sum of costs of all slices, where parameter X can represent the weight matrix, denoted by W :

$$\Phi(X) = \sum_{s=1}^S \Phi_s^A(X) + \sum_{s=1}^S \Phi_s^M(X) \quad (7)$$

Goal is to minimize this cost function ultimately, or in other words, to search for the minimum deviation from target KPIs or better achievement of SLAs. It should be noted that the problem constraint is the positivity and normalization of weights in all cells.

2-2-1-3- Cost Function

For each Key Performance Indicator (KPI), the following criteria have been considered in defining cost functions:

- Cost functions should be aligned with the nature of a service's KPI. For instance, if latency is used as a KPI, the aim is to decrease this value. However, if the average bandwidth KPI is the issue, the value should increase.
- Cost functions should be normalized so that the total cost is not influenced by a single cost function. For example, if one cell has a significant deviation from the Service Level Agreement (SLA), while other cells are close to the SLA, the impact of the other cells should not be observed in the calculations.
- Cost functions should be able to prioritize slices. For example, a slice with excessive traffic should have lower priority.

Cost functions are defined as follows:

$$\Phi_s^A(X) = \left(1 - As(W) \cdot \frac{ls}{\hat{A}_s} \right)^2 \cdot H(1 - As(W) \cdot \frac{ls}{\hat{A}_s}) \quad (8)$$

$$\Phi_s^M(X) = \left(1 - Ms(W) \cdot \frac{ls}{\hat{M}_s} \right)^2 \cdot H(1 - Ms(W) \cdot \frac{ls}{\hat{M}_s}) \quad (9)$$

$$Is = \max \left(\frac{Ls}{\hat{L}_s} \right) \quad (10)$$

The step function $H(.)$ and the extra load index ls . The result of equation (10), which is used in equations (2) and (11), can prioritize different slices based on the defined load constraints in SLA and the instantaneous load imposed on the network by the slice. To clarify the expressions provided and considering the definition of the cost function, it can be understood that if the average bandwidth or the minimum calculated bandwidth from the network exceeds the value defined in the SLA, the step function $H(.)$ will be zero according to the relationship, and that slice will not play a role in calculating the final cost function. On the other hand, the closer this calculated value is to the agreed value, the less numerical impact it has on the final cost function. Now, if this calculated value falls below the SLA, with

the help of optimization algorithms, the best weight for each slice is selected to bring the obtained *KPI* closer to the *SLA*.

3- Proposed method

This work adopts an iterative optimization algorithm for SLA-driven radio resource management in 5G network slicing. The algorithm defines a cost function capturing deviations from target KPIs for each slice, incorporating slice priority and traffic load considerations. Control parameters (slice resource weights) are updated iteratively and locally at each cell, guided by KPI deviation vectors to promptly address SLA violations. High-priority slices receive accelerated adjustment steps to ensure rapid SLA compliance. The method advances prior global update approaches by focusing on local, per-cell updates, thus enhancing responsiveness and reducing negative impact on non-congested slices. This algorithm is an iterative gradient-like optimization of slice weights with:

- Piecewise quadratic cost functions focused on SLA violations.
- Asymmetric step updates prioritizing SLA recovery for critical slices.
- Local cell-based weight adjustments using KPI deviation vectors.
- Normalization constraints ensuring weights sum to 1 per cell.

It balances efficient resource allocation and SLA fulfillment by continuously adapting slice resource shares based on real-time network performance.

RAN slicing deals with efficient sharing of radio resources, namely time and frequency, among subscribers. Slicing RAN is more challenging than slicing CN because RAN slicing involves the distribution of radio resources that are inherently fluctuating and their expansion is more difficult due to the random and dynamic nature of wireless environments, different capacities of radio resources, and system-level abstractions. This study has presented a flexible and adjustable framework for RAN slicing where various requirements of slices are simultaneously considered, and slicing management algorithms adjust various parameters to control the management of radio resources to meet the service-level agreements of the slices. All of these algorithms and requirements have been thoroughly discussed in the previous section. With an overview of the adaptation methods for mapping layer discussed in previous studies, the aim here is to propose a method that introduces SLA profiles and prioritization to have better organization for updating control parameters in the presence of anomalies.

3-2. Proposed Framework

The goal of a RRM (Radio Resource Management) system is to be aware of the load and adjust control parameters for different slices in various cells in a way that network KPIs (Key Performance Indicators) are maintained at a level that does not violate any of the slice SLAs (Service Level Agreements). The proposed model is similar to the system in the referenced paper [18], therefore, we first examine the simulation environment and system details used in the mentioned research and then explain the applied changes.

3-2-1. Proposed System Model

It is assumed that the entire service environment is a cellular network with 7 adjacent cells (similar to Fig. 1). Users enter the network according to a Poisson distribution. The user entry points follow a two-dimensional Gaussian distribution (as shown in Fig. 2) with a density factor $1/\sigma$ that approaches zero. The distribution becomes uniform as this factor approaches zero, and approaches Gaussian as it moves away from zero.

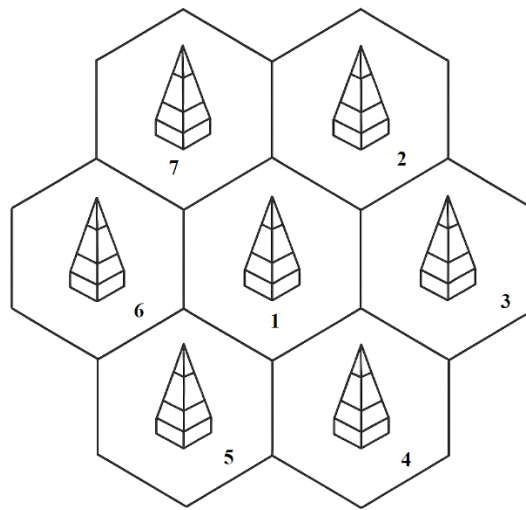


Fig. 1: Overview of the General Service Provision Area [17].

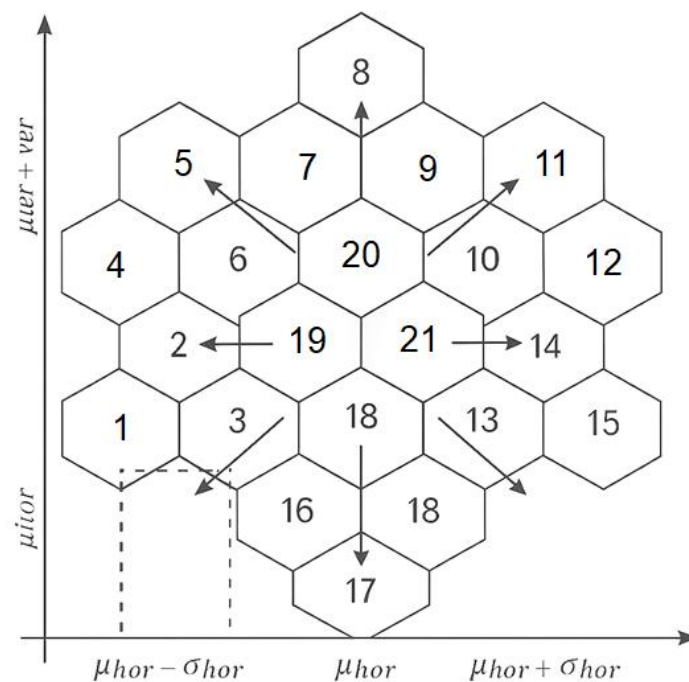


Fig. 2: Gaussian Distribution of Users (for ease of work $\mu_{hor} = \mu_{ver} = \mu$ and $\sigma_{hor}^2 = \sigma_{ver}^2 = \sigma^2$ will be perfectly symmetric due to the two-dimensional Gaussian distribution [17].

The overall environment is composed of three introduced slices: CBR, MBR, and BE. However, the introduced framework here is not limited and allows for the addition of various slices with different KPIs or various RRM mechanisms. The specifications of each of these slices' thresholds are presented in Table 3. The values related to agreed-upon levels and the volume of downloaded files in the FTP model are selected in a way that, before and prior to any abnormalities occur, the network can meet all SLAs without the need to change control parameters.

Table 3: Specifications of Different Slices

Slicing model	BE	CBR	MBR
Desired service	eMBB	URLLC	eMBB and URLLC
Control parameters	1. The weight assigned by the mapping layer to determine the bandwidth share	1. Entry permit limit	1. The weight assigned by the mapping layer to determine the bandwidth share. 2. Entry permit limit
Traffic parameters	1. Entry rate 2. File volume	1. Entry rate 2. File volume 3. Constant bit rate	1. Entry rate 2. File volume 3. Minimum bit rate
KPIs and values of SLA	1. Average Transmittance = 12 Mbps 2. 5% Transmittance=97.5% 3. Dismissal rate from Network = 97.5	1. Acceptance rate =97.5%	1. Acceptance rate =97.5% 2. Average Transmittance = 12 Mbps

The system model is designed such that users enter the network randomly and are assigned to one of the sectors based on an agreement. Depending on their conditions, they download a 15MB file and then exit the network. If a user remains in sector BE for more than 7.5 seconds, they will be removed from it. Network control parameters, similar to previous research, are updated every 60 seconds based on the received Key Performance Indicators (KPI) conditions. The total available bandwidth is 90MHz, which is used in calculating the momentary bit rate of users in the sectors using the Shannon formula. The radius of each cell is considered to be 1km. Typically, macro cells have a radius ranging from 400 meters to 2 kilometers, while macro cells have a radius of 1 to 30 kilometers, depending on factors such as population density [22]. The details such as KPIs, SLAs, user distribution, and service environment specifications are provided in Table 4. To evaluate the performance of different schemes, anomalies are introduced into the network and observe how each method reacts to these anomalies. Anomalies could refer to increased traffic load or congestion factors. A scheme that can achieve the highest KPI in both traffic load and congestion factor outperforms other methods. An additional idea to previous research is introduced following the layer mapping adaptation method presented in [18], in the form of a new method called "Method 5" added to the research: Method 5: Centralized adaptation, with the ability to adjust control parameters for each cell separately along with prioritizing the updating of control parameters (resources) for slices based on the local KPI deviation vector in method 4, similar to method 3, a central entity called the mapping layer receives the overall network reports. The main difference between method 3 and method 4 is that in method 3, the deviation criterion was a general deviation vector, and this layer only had the ability to change overall coefficients, meaning if it felt that the entire network deviated from one or more KPIs (the list of these KPIs is provided in table 4), the only action it could take was to update all coefficients and couldn't change the coefficients of just one cell. For example, if the reason for the abnormality of general conditions is cell number 1 and other cells operate normally, changing the overall coefficients through the mapping layer would result in the coefficients of all cells being altered. This fact would cause, for instance, if the KPI related to slicing CBR in cell 1 deviated, an increase in the threshold related to slicing CBR in all cells by the mapping layer would reduce the threshold for slicing MBR in other cells, leading to the endangerment of MBR slicing KPIs,

as it is evident that with an increase in the share of slicing CBR in one cell, the share of the other two slicing types decreases in proportion. In other words, enhancing the conditions of one slicing type may jeopardize the conditions of other slicing types. As mentioned in the previous section, in method 4, by changing the deviation vector, from the general deviation vector to the Hadamard product of local and general deviation vector ($V^t \odot V_c^t$), and enhancing the accessibility of the mapping layer to locally modify parameters for each cell, an attempt was made to improve the critical situation. The issue with this method is that the mapping layer does not seek to adjust control parameters until a general problem and deviation arise for the network. For example, by increasing the density factor, cell number 1 encounters issues, causing a decrease in the CBR slice acceptance rate in that cell. However, since the KPI status is good in other cells (due to a constant number of inputs but increased density factor leading to more users entering cell 1 and reducing the load on other cells in this regard), the mapping layer does not come into play, creating a Single point of failure in cell 1 where users' QoE significantly decreases. This drawback also applies to method 3. The proposed method suggests that the deviation vector be the same as the local deviation vector, and in the event of any deviation in any of the cells, since the mapping layer has the ability to change control parameters locally, it can take steps to improve conditions locally. In this case, whenever a deviation occurs in any of the cell slices, at that moment, the mapping layer comes into play. Another issue with the previous method was that, in each time interval, the updating coefficient, denoted as δ , was constant and equal to 1.0. This coefficient's stability causes sensitive KPIs such as A_{CBR} to consume a lot of time and resources to improve their conditions, leading other KPIs to gradually deteriorate. As conditions for A_{CBR} improve, other KPIs like A_{MBR} worsen, and this trend continues until the end of the simulation period. For solving the above problem, in the first stage, the values of increase coefficients are updated and decreased separately. In other words, the coefficient of increasing the CBR slicing threshold is considered twice the coefficient for decreasing it. In other words, the updating coefficient for the CBR slicing threshold in a cell changes by two units δ if an increase is needed and by one unit δ if a decrease is required, so that compared to the constant coefficient, the CBR slicing threshold has more opportunity to improve slicing conditions. For example, under certain circumstances, the CBR slicing threshold is 0.2, and this slicing point cannot meet its KPI which is the acceptance level. In this method, instead of adding 1.0 to this threshold, 2.0 is added so that an appropriate offset and less time are needed to ensure the KPI, because adding one unit may result in some users not having access permission to the slicing point again. On the other hand, if the KPI conditions related to the CBR slicing point are appropriate and the mapping layer allows for improving the MBR slicing conditions and there is a need to decrease the CBR slicing threshold percentage in the cell, then the same coefficient of 1.0 is reduced from the threshold so that KPI does not experience a significant and momentary decline.

On one hand, it is evident that in the network, there is a priority for flows, and CBR, MBR, and BE flows have high priority in that order. In other words, as long as the conditions for CBR flows are critical, the congestion control layer will not improve conditions for other flows. To clarify the above explanations, a review of some necessary concepts is required.

3-2-1-1. Service Classification and Prioritization

CoS, (Class of Service), is a method for managing traffic in a network that groups similar types of traffic (such as high-quality video streaming, augmented reality, or virtual reality applications that require high data volume) together and considers each type as a class with its own service priority level. On the other hand, the combination of software-defined networking and virtualized networks enables reducing the complexity of network management. In particular, the ability to manage multiple logical separated virtual networks (slices) plays a significant role in Internet of Things applications, as these applications need to coexist with the management of various traffic streams with specific QoS conditions. Prioritizing slices is not only related to the initial moment of user entry but is a process that should be considered throughout the entire slice management period. Based on the available network capabilities, RAN, transmission network, core network, security aspects, and functional capabilities,

considerations have been made for different levels of network slices to conform to network deployment with the maximum possible policies [23]. For example, in Huawei's research [24], as illustrated in Figs 3, a total of five slicing levels have been provided to meet the three critical 5G life-critical needs. It should be noted that there are some sections that are of high importance, and the necessity of having a mechanism to ensure their services is strongly felt.

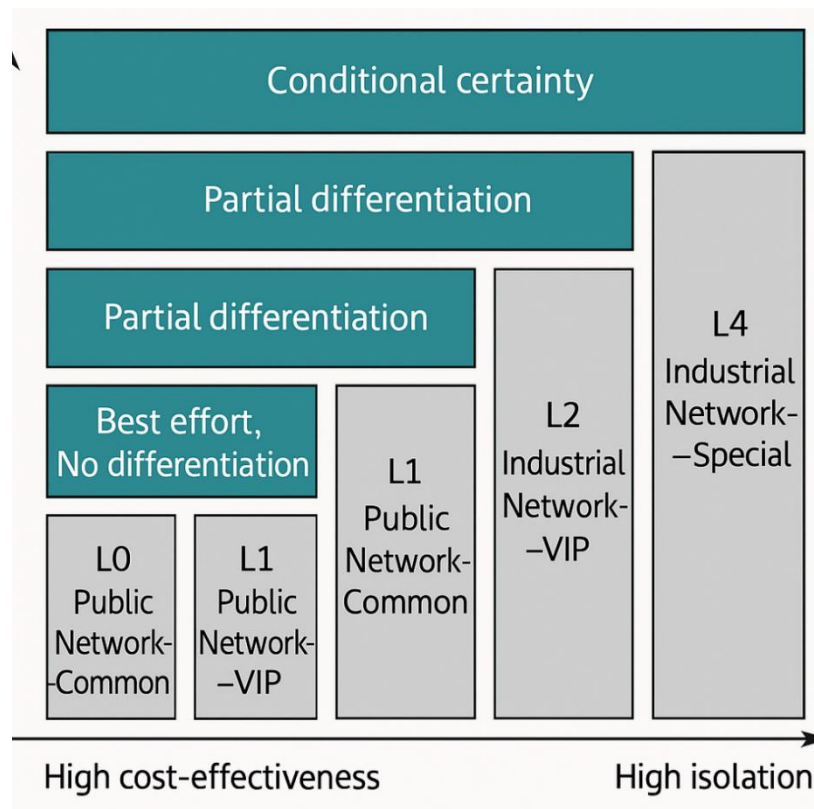


Fig. 3: Introduction of feature planes for slices presented by Huawei for various 5G applications.

Table 4: Features and requirements of 5 different slices introduced by Huawei.

Slice Level	Network Type	Level Classification	Definition	Resource Customization/Resource Isolation	Resource Customization Security	Service Experience O&M	Service Experience Customized Service
L0	Public network	Common	Built on 5G public network infrastructure, with no special requirements	Complete sharing	Basic security	None	Default
L1		VIP	Built on 5G public network infrastructure, with tailored requirements	Complete sharing (or partially exclusive)	eMBB enhanced security	None	Customization
			Built on 5G industry				

L2	Industrial network	Common	network infrastructure, with value-added services	Complete sharing (or partially exclusive)	Service feature security	Visualization	Customization
L3		VIP	Built on 5G industry network infrastructure, with specific resources exclusively used and advanced services provided	Partially exclusive	High-level service feature security	Manageability	Customization
L4		Special	Built on 5G industry private network infrastructure, with all resources exclusively used and reliability services provided	Completely independent	High-level security	Manageability	Customization

4- Results and Analysis

In order to evaluate the proposed designs and algorithms, first, the simulation details are described. It is assumed that there are three types of bearers in the network: BE, CBR, and MBR. Along with these bearers, the traffic load, user distribution, and control parameters are listed in Table 2. To assess the performance of the proposed method compared to the method introduced in [18], anomalies are introduced into the network and observed to see how each method can react to these anomalies. These anomalies can include an increase in traffic load (the number of users entering at a given moment) or the congestion factor. A design that can meet all KPIs at higher levels of traffic load and congestion is superior to another method, in other words, a method that is closer to meeting SLA under difficult conditions is better.

Table 5: Simulation Parameters

Table of Simulation Parameters	
The size of the file to download	15Mb
Time to update control parameters	1 min
The step of updating the control parameters	Variable (0.1 and 0.2)
Running time for each point	30min

Threshold of the user's dismissal time in the BE slicing	7.5 sec
Total bandwidth	90 MHz
The number of cells in the service area	7
Cell radius	1 Km
Antenna model	Omni-directional
CBR slicing load	2-3-4 [users/s/cell]
MBR slicing load	2-3-4 [users/s/cell]
BE slicing load	7-10-11 [users/s/cell]
Distribution of users in all slices	Uniform to Gaussian
Default weight of mbr slice	0.5
CBR slice threshold	0.4
MBR slice threshold	0.4

4-1. Simulation Results and Analysis

In this section, the proposed method is compared with the method presented in the fourth paper [18] considering different distributions and density coefficients. Generally, in this part, the RAN slice is studied from a system-level perspective, and hypotheses about the physical layer mechanism are considered in a way that the main result of this work is not influenced by these assumptions. It is assumed that instead of allocating physical resource blocks, surplus resources can be allocated to users. In summary, it can be said that the Signal-to-Interference plus Noise Ratio (SINR) ratio depends on the following conditions [25]:

- Distance between ENB (EndNodeB) and UEs (User equipment)
- Objects present between ENB and UE such as buildings, trees, and...
- Interference from other cells (especially channel interference)
- Interference from the same cell (if using multi-user MIMO)
- Antenna gain (increasing it significantly can reduce interference)
- Beam-forming positioning in the scattered channel
- Transmitter power levels
- Number of transmitters

In this section, to calculate the Signal-to-Interference-Plus-Noise Ratio (SINR), we first consider the worst-case scenario where interference always exists from surrounding cells [26]. It is worth mentioning that if for any reason the SINR of a user is very low, on the order of a few tenths of decibels, in order for the desired user to access their preferred service, significant network resources are occupied for a considerable amount of time. To address this issue, the SINR of users in a reasonable range of 10

to 11 decibels (mid cell) is considered, with the aim of maintaining SINR constant throughout the simulation period to observe the impact of the proposed approach.

Table 6: Different SINR levels based on the location in the cell and received signal conditions.

	Condition	SINR(dB)	SINR(dB)	PSRQ(dB)
RF Condition	Excellent	≥ 20	≥ -10	≥ -80
	Good	13 to 20	-10 to -15	80 to 90
	Min Cell	0 to 13	-15 to -20	-90 to 80
	Cell Edge	≤ 10	< -20	≤ -100

In the next stage, in order to simulate longer time intervals (e.g., 30 minutes), the aspect of user mobility is disregarded, hence, SINR values remain constant throughout the download period. The overall system model is such that users belonging to each sector enter the network at random times and locations, intending to download a 15Mb file and then exit the network (traffic model FTP) [27]. The fourth method mentioned is aware of the entire network but only modifies control parameters when not only local deviation occurs but also the corresponding KPI deviates globally. In the fifth method, this update is done locally. Additionally, for CBR slicing, which has the highest priority, the updating coefficient has increased to 0.2 in the increase case and decreased to 0.1 in the time reduction scenario. Furthermore, for improving slice conditions, improvement priority is considered. Consequently, for a mapping layer that has poor status for CBR slicing, other slices are not important, and similarly, for MBR slicing against BE. The overall flowchart of the proposed method is shown in Fig. 5. The Fig. number 6 compares the status of the first cell (central area of service region) in two proposed methods. It should be noted that as we move towards the right of the graph, the density factor increases, and more users enter the first cell in the same distribution. In general, as the density factor increases and the distribution approaches a Gaussian distribution, the status of the central cell becomes more critical, while the status of other cells improves. One of the reasons for the weakness of the fourth method in dealing with anomalies is that as the density coefficient or the number of input users increases, on the one hand, the status of the first cell tends towards deterioration and, on the other hand, the statuses of subsequent cells improve. This is because new users, due to the spatial distribution approaching Gaussian, tend to occupy the first cell more, providing fewer services to other cells. This point leads to the assumption of an appropriate general status from the mapping layer perspective; the overall deviation vector is considered zero, consequently, its inner product with the local deviation vector becomes zero, and this layer does not have a reactive response to this situation of deterioration. In other words, in contrast to the fifth method, the mapping layer's response to improving conditions is much slower in the fourth method.

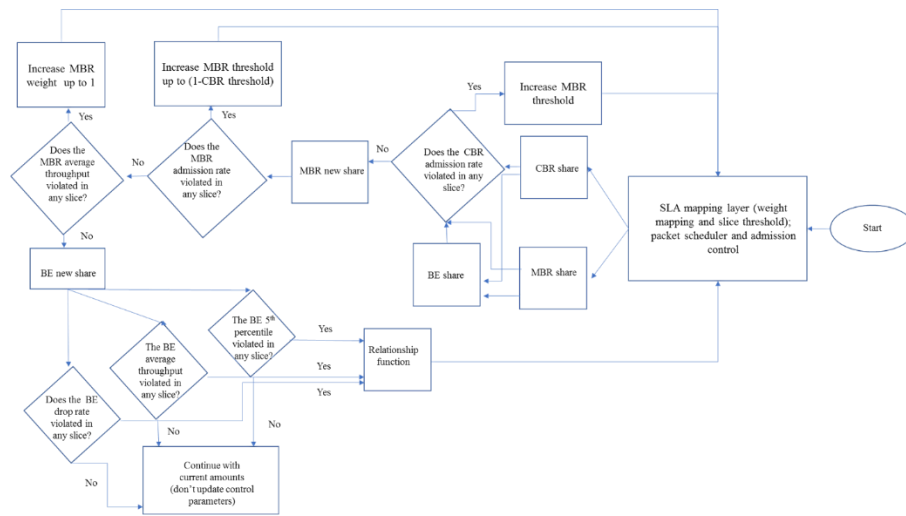


Fig. 5. The flowchart of the proposed method

Since the priority of method 5, or the proposed method, is to achieve KPIs of CBR, MBR, and BE in order, it can be clearly observed that the conditions of CBR and MBR slices have improved, while on the other hand, the conditions of the BE slice have deteriorated slightly. As the simulation runtime increases, the conditions of CBR and MBR slices improve because as time increases, the total number of input users increases and the proportion of a few users who may not receive permission for updates decreases. It is necessary to mention that the BE slice, as its name suggests, makes maximum efforts to provide the best conditions to the user, but it does not guarantee anything, so it is better to use the resources of this slice to provide services to higher priority slices in case of network abnormalities. On the other hand, it is necessary to explain that the responsibility of the mapping layer is to ensure SLA and it does not intervene for values exceeding it; for example, the task of this layer in relation to the CBR threshold is to provide 97.5% input (all these values are listed in table number 4) and if the input percentage falls below this value, it will come into effect. In other words, a deviation must occur for the mapping layer to become aware of it. This specific characteristic can be considered one of the weaknesses of this system, and to address this issue, artificial intelligence methods can be utilized to identify system behavior so that the mapping layer can intervene before a KPI deviation occurs.

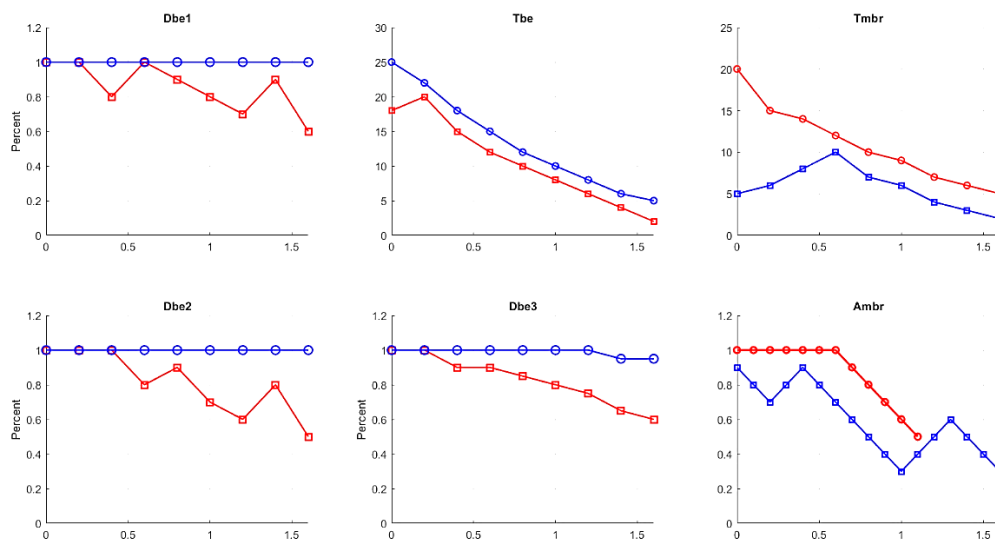


Fig. 6: Results related to the mentioned methods in the presence of different traffic volumes related to the central cell or the first. The graphs are based on the KPI value (these values vary depending on the type of KPI, for example, for *Acbr* in percentage and for *Tmbr* in Mbps) against the congestion factor.

The meaning of (4-5-11) is the number of incoming users at any moment according to the exponential distribution to BE, MBR and CBR sections respectively. Figure 6 compares the condition of the first cell (the center of the service area) in the two proposed methods. As we move to the right side of the graph, the density coefficient increases and more users enter the first cell in the same distribution number. In general, with the increase of the density factor and the distribution approaching the Gaussian distribution, the condition of the central cell deteriorates and the condition of other cells improves. One of the reasons for the weakness of the fourth method in dealing with anomalies is that as the density factor or the number of incoming users increases, on the one hand, the condition of the first cell deteriorates, and on the other hand, the condition of the next cells improves, because new users are approaching Spatial distribution is like Gaussian, most are placed in the first cell and other cells serve fewer users.

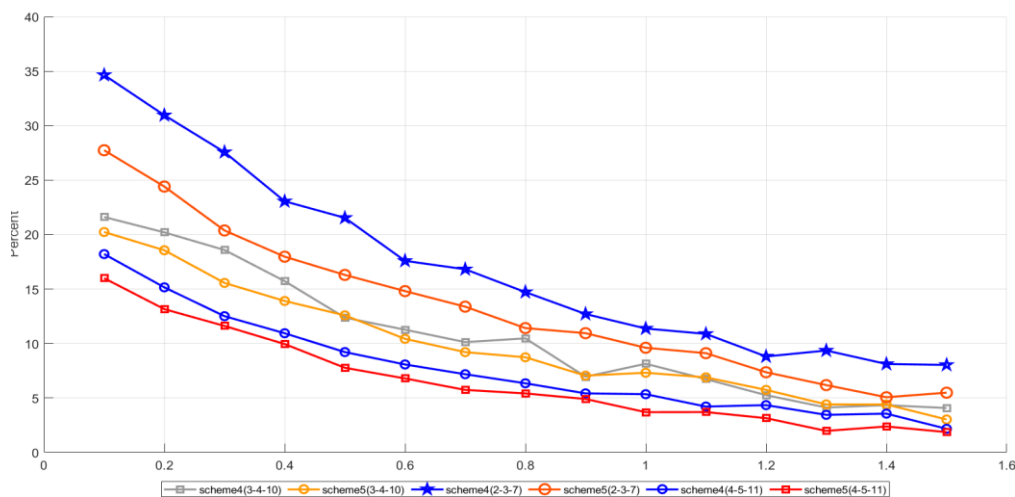


Fig. 7: Results related to the mentioned methods in the presence of different traffic volumes related to the central cell or the first. The graphs are based on the KPI value (these values vary depending on the type of KPI, for example, for *Acbr* in percentage and for *Tmbr* in Mbps) against the congestion factor.

The point that is clearly illustrated in Fig. 7 is that as the number of input users increases at any given moment (indicated by different colors), the graph tends to decrease, with a greater decline as the number increases. Regarding graph Tbe, it is evident that as the traffic load increases, all methods converge to a similar range, indicating that at higher loads, the proposed method performs relatively well compared to the previous method. However, the three lower graphs correspond to the proposed method. Graphs Dbe and Fbe also exhibit a similar pattern. This is because slice BE holds the lowest priority among slices, so it is expected that in comparison to the equal priority state in the fourth method, its condition is worse. The superiority of the proposed method is clearly evident in the *Acbr* graph, where the fourth method shows a significant decrease in traffic load in each of the three traffic instances (represented by different colors in the graph). This indicates that as the distribution approaches a Gaussian distribution and the number of input users increases at any given moment, the proposed method brings about notable stability for CBR slicing. On the other hand, the graph related to *Ambr* also suggests the superiority of the aforementioned method, as it demonstrates a more acceptable performance under harsher conditions. In other words, as we move towards the right side of the graph, the decrease in the graph is less pronounced in the proposed method. In the *Tmbr* graph, similar to the two aforementioned graphs, the fourth method is noticeably inferior to the fifth method, implying that under adverse conditions, it provides a higher bandwidth for users. The output of each of

the distributions is available separately according to different densities, but due to the lengthening of the explanations, it has been omitted.

It should be noted that the oscillatory movement and zigzagging of the graphs are related to the inherent random nature of the two important factors of temporal and spatial distributions of users. In simpler terms, it is true that the traffic load of the 2-3-7 state is significantly less than that of 3-4-10. However, because the timing of user arrivals to the slice follows an exponential distribution, it is possible that at different moments, the number of incoming users for the 2-3-7 traffic load may exceed that of 3-4-10, leading to a localized drop in the bandwidth width or entry rate associated with that distribution slice, causing the graph to take on a zigzag pattern itself, but it is tangible in the overall trend. The same point is valid for the concentration coefficient directing the spatial distribution from uniform to Gaussian - for example, the difference between the states $\delta=0.2$ and $\delta=0.3$ is also accurate.

Fig. 8 compares the average throughput (Acbr) across various traffic loads for the proposed method (Method or Scheme 5) and existing methods 4 from [18]. It visualizes how each method performs in maintaining and improving Acbr as traffic demand increases.

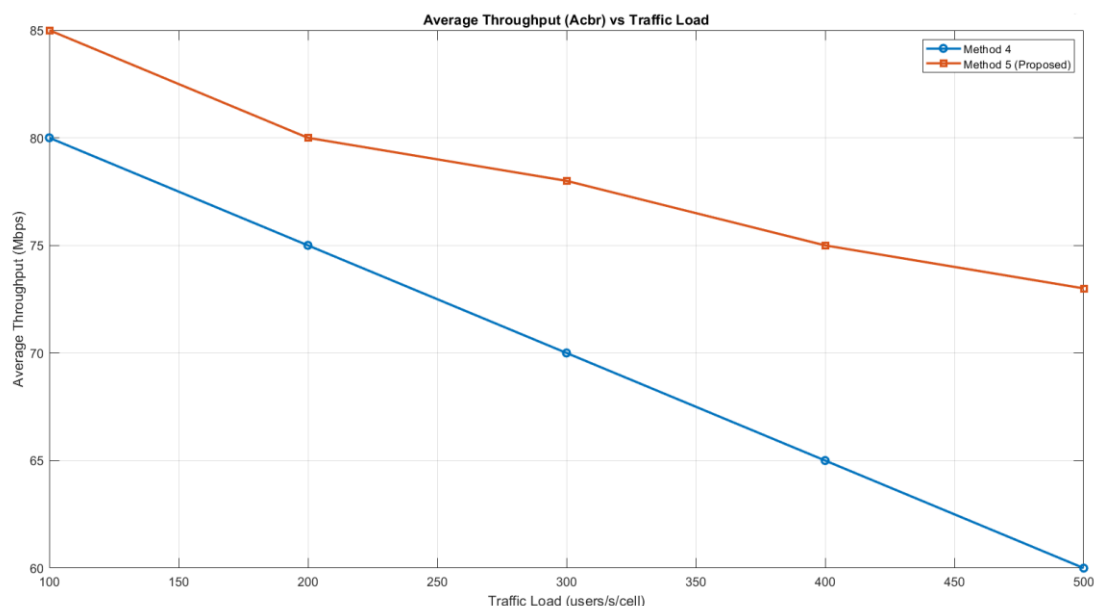


Fig. 8. comparison of the average throughput (Acbr).

4-2. Simulation Environment and Simulation Device

This simulation was executed in a Python programming environment on a system with 16GB RAM and a graphics card with 10GB memory, specifically the Nvidia Quadro K2100M model. Each point of the graphs 8 and 9 was run for 30 minutes, with Key Performance Indicators (KPIs) being constantly available and monitorable.

- Conclusion

It is expected that 5G mobile networks will support flexible and scalable needs; therefore, network resources should be dynamically allocated based on requests. Network slicing, where network resources are packaged according to specific user needs and allocated separately to each, is considered a key model for achieving diversity in needs. The allocation of resources to these slices will clearly involve conflicting or somewhat ambiguous requirements, and how the network addresses these needs is of utmost importance. In this study, a novel admission control mechanism is proposed that can allocate network resources to maximize user satisfaction while ensuring the requirements of the corresponding slices.

Through simulation, it was demonstrated how the proposed method enables improved utilization of radio resources and higher scalability and stability when the number of users in each slice increases. Since users expect a level of service corresponding to their slice, this approach may even enhance the user experience significantly compared to the fourth method, particularly if using metrics related to user experience evaluation.

REFERENCES

- [1] M. Condoluci and T. Mahmoodi, "Softwarization and virtualization in 5G mobile networks: Benefits, trends and challenges," *Comput. Networks*, vol. 146, pp. 65–84, 2018.
- [2] B. Khodapanah, A. Awada, I. Viering, D. Oehmann, M. Simsek, and G. P. Fettweis, "Fulfillment of service level agreements via slice-aware radio resource management in 5G networks," in *2018 IEEE 87th Vehicular Technology Conference (VTC Spring)*, 2018, pp. 1–6.
- [3] GSMA, "Network slicing use case requirements," 2018.
- [4] N. Alliance, "Description of network slicing concept," *NGMN 5G P*, vol. 1, no. 1, 2016.
- [5] 3GPP, "3rd Generation Partnership Project; Technical Specification Group Services and System Aspects; Management and orchestration; Concepts, use cases and requirements (Release 15)," 2019.
- [6] B. Khodapanah, A. Awada, I. Viering, J. Francis, M. Simsek, and G. P. Fettweis, "Radio resource management in context of network slicing: What is missing in existing mechanisms?" in *2019 IEEE Wireless Communications and Networking Conference (WCNC)*, 2019, pp. 1–7.
- [7] I. Da Silva *et al.*, "Impact of network slicing on 5G Radio Access Networks," in *2016 European conference on networks and communications (EuCNC)*, 2016, pp. 153–157.
- [8] 3rd Generation Partnership Project (3GPP), "Deliverable D2.4 Final Overall 5G RAN Design," 2017.
- [9] Khalek AA, Al-Kanj L, Dawy Z, Turkiyyah G. "Optimization models and algorithms for joint uplink/downlink UMTS radio network planning with SIR-based power control," [ieeexplore.ieee.org](https://ieeexplore.ieee.org/abstract/document/5739544/), Accessed: Aug. 02, 2021. [Online]. Available: <https://ieeexplore.ieee.org/abstract/document/5739544/>.
- [10] M. Richart, J. Baliosian, J. Serrat, and J.-L. Gorricho, "Resource slicing in virtual wireless networks: A survey," *IEEE Trans. Netw. Serv. Manag.*, vol. 13, no. 3, pp. 462–476, 2016.
- [11] C. Liang and F. R. Yu, "Wireless network virtualization: A survey, some research issues and challenges," *IEEE Commun. Surv. Tutorials*, vol. 17, no. 1, pp. 358–380, 2014.
- [12] Pérez-Romero, Jordi, Oriol Sallent, Ramon Ferrús, and Ramón Agustí. "Self-optimized admission control for multitenant radio access networks." In *2017 IEEE 28th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC)*, pp. 1-5. IEEE, 2017.
- [13] F. 2015 NGMN Alliance, "NGMN 5G White Paper," Tech. Rep., "No Title," 2015.
- [14] Ksentini, Adlen, and Navid Nikaein. "Toward enforcing network slicing on RAN: Flexibility and resources abstraction." *IEEE Communications Magazine* 55, no. 6 (2017): 102-108.
- [15] M. Jiang, M. Condoluci, and T. Mahmoodi, "Network slicing management & prioritization in 5G mobile systems." Accessed: Apr. 09, 2021. [Online]. Available: <https://ieeexplore.ieee.org/abstract/document/7499297/>.
- [16] Khodapanah, Behnam, Ahmad Awada, Ingo Viering, Andre Noll Barreto, Meryem Simsek, and Gerhard Fettweis. "Slice management in radio access network via deep reinforcement learning." In *2020 IEEE 91st Vehicular Technology Conference (VTC2020-Spring)*, pp. 1-6. IEEE, 2020.
- [17] B. Khodapanah and G. S. Member, "Framework for Slice-Aware Radio Resource Management Utilizing Artificial Neural Networks," vol. 8, 2020, doi: 10.1109/ACCESS.2020.3026164.

- [18] B. Khodapanah, A. Awada, I. Viering, A. N. Barreto, M. Simsek, and G. Fettweis, "Slice Management in Radio Access Network via Iterative Adaptation," *ICC 2019 - 2019 IEEE Int. Conf. Commun.*, pp. 1–7, 2019.
- [19] D. Zhang, ... Z. C.-2016 I. 83rd V., and undefined 2016, "Reverse combinatorial auction-based resource allocation in heterogeneous software defined network with infrastructure sharing," *ieeexplore.ieee.org*, Accessed: Apr. 11, 2021. [Online]. Available: <https://ieeexplore.ieee.org/abstract/document/7504455/>.
- [20] M. Jiang and M. Condoluci, "Network slicing in 5G: An Auction-Based Model." Accessed: Apr. 11, 2021. [Online]. Available: <https://ieeexplore.ieee.org/abstract/document/7996490/>.
- [21] O. Narmanlioglu, E. Zeydan, S. A.-I. Access, and undefined 2018, "Service-aware multi-resource allocation in software-defined next generation cellular networks," *ieeexplore.ieee.org*, Accessed: Apr. 11, 2021. [Online]. Available: <https://ieeexplore.ieee.org/abstract/document/8323373/>.
- [22] Jean-Paul Linnartz, "web site." <http://www.wirelesscommunication.nl/reference/chaptr04/cellplan/cellsize.htm>.
- [23] China Electric Power Research Institute, Tencent, China Sports Media, Gree, Digital Domain Group, and AsiaInfo Technologies, "Categories and Service Levels of Network Slicing White Paper," no. March, 2020.
- [24] China Electric Power Research Institute, Tencent, China Sports Media, Gree, Digital Domain Group, and AsiaInfo Technologies, "huawei." <https://www-file.huawei.com/-/media/corporate/pdf/news/categories-slice--white-paper-en.pdf?la=en>.
- [25] Naser Khatte, Emil Björnson, Naveed Iqbal, Jiankang Zhang "www.researchgate.net." <https://www.researchgate.net/post/Hi-dears-im-working-on-SINR-issue-on-LTE-network-and-im-looking-for-all-parameters-which-can-degrade-SINR-value-has-anybody-information-about-it>.
- [26] M. Castañeda, M. T. Ivrlač, J. A. Nossek, and A. Klein, "The 18th Annual IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC'07) ON DOWNLINK INTERCELL INTERFERENCE IN A CELLULAR SYSTEM." Accessed: May 01, 2021. [Online]. Available: <https://ieeexplore.ieee.org/abstract/document/4394052/>.
- [27] 3GPP, Evolved Universal Terrestrial Radio Access (E-UTRA); User Equipment (UE) radio transmission and reception (3GPP TS 36.101 version 14.3.0 Release 14) "E_utra," 2017.
- [28] Suh, Kyungjoo, Sunwoo Kim, Yongjun Ahn, Seungnyun Kim, Hyungyu Ju, and Byonghyo Shim. "Deep reinforcement learning-based network slicing for beyond 5G." *IEEE Access* 10 (2022): 7384-7395.
- [29] Yan, Dandan, Benjamin K. Ng, Wei Ke, and Chan-Tong Lam. "Deep Reinforcement Learning Based Resource Allocation for Network Slicing With Massive MIMO." *IEEE Access* (2023).