Assessment of Key 5G RAN Features for Integrated Services Provisioning in a Smart City Environment

Ayushi Kandoi

TU Delft

Assessment of Key 5G RAN Features for Integrated Services Provisioning in a Smart City Environment

by

Ayushi Kandoi

in partial fulfilment of the requirements for the degree of

Master of Science
in Electrical Engineering
Track Wireless Communication and Sensing

at the Delft University of Technology,
to be defended publicly on Tuesday, May 31, 2022 at 2:00 PM.

Student number:	5024498	
Project duration:	December 20, 2020 – May 31, 2022	
Thesis committee:	Dr. Remco Litjens, MSc	TU Delft, TNO
	Dr. Georgios Iosifidis	TU Delft
	Maria Raftopoulou	TU Delft, Daily Supervisor

TU Delft

Preface

This thesis marks the culmination of my studies at the Delft University of Technology for the Master of Science degree in the field of Electrical Engineering with the specialization in Wireless Communication and Sensing. The completion and success of this thesis would not have been possible without the timely guidance and constant support from many individuals.

Working on this thesis in the desired area of interest was very enthusiastic and exciting journey for me. I would like to first thank my daily supervisor, Maria Raftopoulou who suggested this research topic and guided me through the whole research process. Without Maria's support, guidance, feedback and constant motivation I would not have been able to achieve the success of this thesis. I would like to specially appreciate her extra effort beyond supervision every week and sometimes over the weekends. Secondly, I would like to express my immense gratitude towards Remco Litjens, my supervisor at TU Delft. His teachings have motivated me to follow this path towards the Radio Access Networks. His expertise and knowledge has provided me with timely guidance and critical feedbacks which have immensely improved the quality of my research. Also, I would like to thank both my supervisors to help me improve my scientific writing skills with their valuable feedback review rounds. Apart from that, I would also like to appreciate their patience and understanding during the tough times with my health and sufferings.

This study's research investigation scenario has been decided in collaboration with KPN. I would like to show my gratitude towards Frank Mertz, Eric Smeitink and Ramin Hekmat at KPN who helped me find the most suitable and realistic investigation scenario for the research and also for providing their feedback from industrial point of view. I am thankful for their valuable time they have invested for this research.

Next, I would like to show my respect and gratitude towards almighty, my parents Mr. Govind Garg and Mrs. Shweta Garg, my brother Keshav Kandoi and my whole family for their constant motivation, concern and immense love. I would also like to thank my friends Viswanath Das, Samanvya Singh, Sandeep Patil, my roommates and classmates in Delft for always supporting and helping me during this research period. Lastly, I would like to thank the faculty of Electrical Engineering for their valuable teachings and the university support centres for helping me with information in the whole master program.

Ayushi Kandoi
The Hague, May 2022

Abstract

The Fifth Generation (5G) network is expected to support three main service categories namely enhanced Mobile BroadBand (eMBB), Ultra-Reliable and Low-Latency Communications (URLLC) and massive Machine Type Communications (mMTC), where each service group has a different Quality of Service (QoS) requirements i.e. the eMBB service group has a high throughput requirement, the URLLC service group require very low-latency and highly reliable transmissions and the mMTC service group do not have a strict performance requirement but have massive connection of devices in the network.

5G enables many vertical domains with these three service categories, such as, smart cities. In a smart city environment, there are applications from all three service categories such as massive connectivity of the sensors for waste managements or monitoring environmental conditions, video surveillance along the city streets and many more. This study addresses the problem of managing applications from the three service categories on the same physical network infrastructure at the Radio Access Network (RAN). Two different scenarios, one with and one without an emergency incident, are considered to find the impact of the incident in the network.

In 5G networks, many new features are introduced such as, flexible numerology, mini-slot based scheduling, BandWidth Parts (BWPs) and RAN slicing. The key objective of this study is to assess the 5G RAN features in terms of achieving the performance requirements of the considered applications, simultaneously. To do so, different RAN configurations are modelled where, a RAN configuration consists of one or multiple RAN features. The evaluation is done by simulating the different possible RAN configurations. The simulations are performed using an existing 5G system-level simulator which is substantially upgraded with the 5G RAN features and is modified to the considered smart city urban macro-cellular environment and with the considered traffic models for each considered application.

To evaluate the performance of each considered application, different performance metrics are defined based on the application requirements. The benefits and/or losses of different RAN features are found and then different RAN configurations are considered with the combinations of RAN features based on the evaluation of each RAN feature. For all different RAN configurations with combination of features, the performance metrics are evaluated and compared with each other to determine the best-performing configurations for the smart city environment, for the scenarios with and without an incident.

Abbreviations

3G Third-Generation.

3GPP Third Generation Partnership Project.

4G Fourth-Generation.

5G Fifth-Generation.

ACK ACKnowledgement.

AI Artificial Intelligence.

AR Augmented Reality.

AWGN Additive White Gaussian Noise.

BLER Block Error Rate.

BS Base Station.

BWP BandWidth Part.

CCTV Closed Circuit TeleVision.

CDF Cummulative Distribution Function.

CI Confidence Interval.

CQI Channel Quality Indicator.

CSI Channel State Information.

DL Downlink.

EDF Earliest Deadline First.

eMBB enhanced Mobile BroadBand.

FDD Frequency Division Duplexing.

FR Frequency Range.

FTP File Transfer Protocol.

gNB gNodeB.

GP Guard Period.

IMT International Mobile Telecommunications.

INI Inter Numerology Interference.

IoT Internet of Things.

ISD Inter Site Distance.

ITU International Telecommunication Union.

KPI Key Performance Indicator.

LOS Line Of Sight.

M-LWDF Modified-Largest Weighted Delay First.

MCS Modulation and Coding Scheme.

MI Mutual Information.

MI-ESM Mutual Information-Effective SINR Mapping.

mIoT massive Internet of Things.

ML Machine Learning.

mMTC massive Machine Type Communications.

MR Maximum Rate.

NACK Negative ACKnowledgement.

NGMN Next Generation Mobile Networks.

NLOS Non-Line Of Sight.

OFDM Orthogonal Frequency Division Multiplexing.

OLLA Outer-Loop Link Adaptation.

PDCCH Physical Downlink Control Channel.

PDF Probability Density Function.

PDSCH Physical Downlink Shared Channel.

PF Proportional Fair.

PL Path Loss.

PRB Physical Resource Block.

PUCCH Physical Uplink Control Channel.

PUSCH Physical Uplink Shared Channel.

QoS Quality of Service.

RAN Radio Access Network.

S Special slot.

SCS Sub-Carrier Spacing.

SINR Signal-to-Interference-plus-Noise Ratio.

SLA Service Level Agreement.

SR Scheduling Request.

SRS Sounding Reference Signal.

TBS Transport Block Size.

TDD Time Division Duplexing.

UE User Equipment.

UL Uplink.

URLLC Ultra-Reliable and Low-Latency Communications.

V2X Vehicle to Everything.

VR Virtual Reality.

Contents

List of Tables

List of Figures

1 Introduction

This chapter gives the background of the research topic and provides the objectives, approach and outline of the study. Section 1.1 explains the expectations from the 5G network and gives a high-level description of the scenario under investigation and its applications. The requirements of the considered applications of the reference scenario are given in Section 1.2. Further, Section 1.3 provide a high-level explanation of 5G RAN features. Section 1.4 explains the challenges and the motivation behind the study, while Section 1.5 provides the related literature review. Based on the motivation and the related work, the objectives of the study are presented in Section 1.6. Finally, Section 1.7 presents the approach taken to achieve the objectives of the study and the outline of the work.

1.1. 5G Technology and Smart city

The 3G and 4G mobile networks were expected to provide high data transfer rates and were very successful, but in comparison with 5G, these data rates are considered low. The 5G network is envisioned to play a key role in incorporating advanced digital technologies in various sectors such as automotive and transport, health care, industry 4.0 and many more [1]. Together with providing a high peak data rate (up to 20 Gbps), 5G is expected to support data transmission with low-latency of around one ms and handle massive connectivity (up to 10^6 devices/km^2) [1]. Ortiz et al. [1] conducted a survey that projects 10-100 times growth in global International Mobile Telecommunications (IMT) traffic from 2020 to 2030. 5G network refers to the IMT requirements issued by International Telecommunication Union (ITU). Ortiz et al. [1] also present an enormous number of new applications that are expected to be supported by 5G. Based on their service requirements, these applications are grouped into three generic categories, as listed below and shown in Figure 1.1:

- **enhanced Mobile BroadBand (eMBB):** Applications that require high throughput. An example application is in-vehicle infotainment services, such as video streaming [2].
- **Ultra-Reliable and Low-Latency Communications (URLLC):** Applications that have stringent latency and reliability requirements. An example application is collaborative robots for industry automation, as it needs fast and reliable transmission of instructions to robots for safe human-robot interaction [3].
- **massive Machine-Type Communications (mMTC):** Applications that do not have a strict performance requirement but have a massive number of device connections that occasionally transmit small data packets, such as sensors monitoring environmental conditions and sensors for smart traffic light system [2].

With the potential of 5G in supporting these three categories of services, 5G enables several vertical domains such as automotive, education, agriculture, logistics and many more to operate more efficiently and emerge smarter digitally. In recent years, cities are striving to become smart by using advanced applications like the Internet of Things (IoT), Augmented Reality (AR), Virtual Reality (VR) and many other applications. A smart city is a complex ecosystem that uses information and communication technologies for the advancement of various sectors [4]. This complex ecosystem aims to provide a safer, cleaner, and more economic environment for improving the quality of life of the citizens [5]. An example of an application in smart cities is video surveillance using Closed-Circuit Television (CCTV) cameras around the city streets and crime-prone areas to improve public

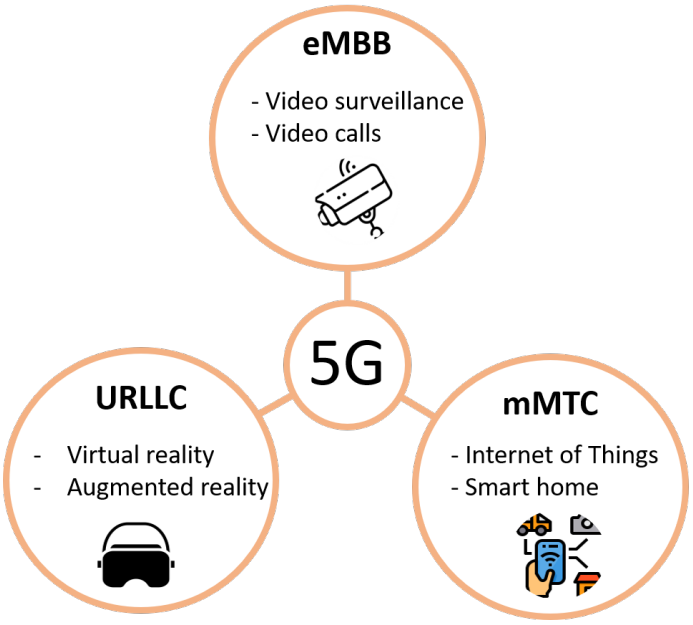

Figure 1.1: The three service categories with example applications.

safety [2]. Another example is the deployment of sensors to control the streetlights for reducing energy consumption [2]. Smart city is an environment with multi-service requirements which can be grouped into the three above-mentioned service categories (eMBB, URLLC and mMTC). Therefore, the smart city is considered as the reference investigation scenario for this study.

In a smart city, there are possibilities of having emergency incidents like road accidents or fire emergencies which require emergency services like control of traffic light system for fast transportation of ambulance, use of body cameras and AR glasses for conducting rescue missions. Therefore, this study takes into account two scenarios: one without any incident and other with an incident to investigate the impact of the incident-based applications on the network. An incident of fire emergency at a building is considered in this study. The considered applications for both normal and incident scenarios and their service requirements are explained in the next section.

1.2. Applications and Requirements

In this study, we consider smart city's normal and incident scenario applications from each of the three service categories described in Section 1.1. The chosen applications and their requirements are described below and summarized in Table 1.1.

- **Normal scenario applications:**
  - **Broadband access everywhere:** The connected smart city requires broadband access to be available everywhere. Next Generation Mobile Networks (NGMN) defined the throughput service requirement of 50 Mbps for Downlink (DL) [6] and 12.5 Mbps Uplink (UL) throughput requirement is considered in this study.
  - **VR:** VR can have various applications in smart cities, such as VR gaming sessions at city malls, sports centers or at home which improves resident's entertainment experience, VR shopping at clothing store and social VR for remote interaction with people. In this study, we only consider VR gaming sessions that constitute continuous visual content with a 360-degree view and 8K resolution, displayed on the VR headset based on change in field of view of the user. The motion feedback of the user needs to be in sync with the

video content arriving at the headset for a high-quality experience. The DL transmission of video packets arriving at the VR headset requires to be delivered within 10 ms latency budget at the RAN. Similarly, the UL transmission of motion feedback data of the user also has a latency requirement of 10 ms for a stable and high-quality experience. The maximum acceptable packet loss rate for both DL **VR video content** and UL **VR motion feedback data** is 4% as specified in [7]. A packet which fails to arrive within the latency budget is considered lost.

- **Video surveillance:** With the emerging efforts towards public safety, video surveillance is commonly used by security officials to improve the safety of citizens, especially in crime-prone areas and around the city streets. The network operator is expected to provide an average UL throughput of 25 Mbps for each CCTV camera used for surveillance purposes [6].
- **Sensors:** The city is connected with the use of sensors that helps the city run more efficiently and improve the quality of living of the citizens. We classify sensors into two groups based on the nature of their measurements:
  - **Sensors for non-risk-sensitive measurements:** Sensors performing non-risk-sensitive measurements such as monitoring environmental conditions like air quality, controlling streetlights and managing the wastes around the city. These sensors occasionally send small amounts of data that do not have a stringent performance requirement, but are massive in number [6].
  - **Sensors for risk-sensitive measurements:** Sensors performing risk-sensitive measurements such as smoke sensors in various buildings, shopping malls and houses around the city. Similar to sensors measuring non-risk-sensitive measurements, these sensors occasionally send small amounts of data with no stringent performance requirement [6].

- **Incident scenario applications:**
  - **Sensors sending emergency messages:** When the sensors performing risk-sensitive measurements detect an abnormality at a certain location, the sensors at that location send urgent messages to the concerned authorities to indicate an unexpected incident. An incident of a fire emergency in a building is considered in this study. The sensors deployed in the incident area are classified as **sensors sending emergency messages** and they have a latency requirement of 30 ms with a maximum packet loss rate of 0.1% [8].
  - **Body camera:** Body cameras are used by the security officials at the incident sites for surveillance of the surrounding incident area, and they stream continuous live video feeds to the control center. An average throughput of 25 Mbps is expected for an undisturbed live-streaming from the incident site [6].
  - **AR:** When there is an incident of a fire emergency in a building, the use of AR glasses for rescue missions is considered in this study. The considered AR application has similar requirements to the previously considered VR gaming application. Specifically, the DL transmission of the video content on the AR glasses and the UL transmission of the motion feedback location of the user need to be in synchronization. Both the DL transmission of the **AR video content** and the UL transmission of the **AR motion feedback data** have a latency requirement of 10 ms with a maximum packet loss rate of 4% [9].

Table 1.1: Applications of smart city and their requirements, as considered in this study.

Application	Scenario type	Service category	Performance requirement		
			Average throughput	Latency	Reliability
Broadband access everywhere	Normal	eMBB	DL: 50 Mbps UL: 12.5 Mbps	-	-
VR (Video content)	Normal	URLLC	-	DL: 10 ms	96%
VR (Motion feedback data)	Normal	URLLC	-	UL: 10 ms	96%
Video surveillance (CCTV)	Normal	eMBB	UL: 25 Mbps	-	-
Sensors for risk-sensitive measurements	Normal	mMTC	-	-	-
Sensors for non-risk-sensitive measurements	Normal	mMTC	-	-	-
Sensors sending emergency messages	Incident	URLLC	-	UL: 30 ms	99.9%
Body camera	Incident	eMBB	UL: 25 Mbps	-	-
AR (Video content)	Incident	URLLC	-	DL: 10 ms	96%
AR (Motion feedback data)	Incident	URLLC	-	UL: 10 ms	96%

1.3. 5G Technological features

The high-level description of the standardized 5G technological features which can help in provisioning the multi-service requirements of the smart city environment are presented in this section. The more detailed description for each feature is presented in the next chapter. The RAN can be configured with individual or combinations of 5G RAN features. The different possibilities available to configure the RAN with the new features are presented in Figure 1.2.

In 5G, the concept of *network slicing* is introduced, where each slice can be seen as a virtual network. With network slicing, multiple independent and isolated end-to-end slices can be created on the same physical network infrastructure. Each end-to-end slice is highly customizable to address diverse requirements in terms of performance, cost, availability and security [10]. This study focuses on provisioning diverse QoS requirements with the help of slicing. Services with similar QoS requirements are grouped in one slice, and each slice can be customized to fulfill the slice-specific service requirements. Hence, network slicing is considered a key enabler feature in 5G to manage the network that serves applications with diverse service requirements. Software-defined network and network function virtualization are the key technologies that enable network slicing in 5G [11].

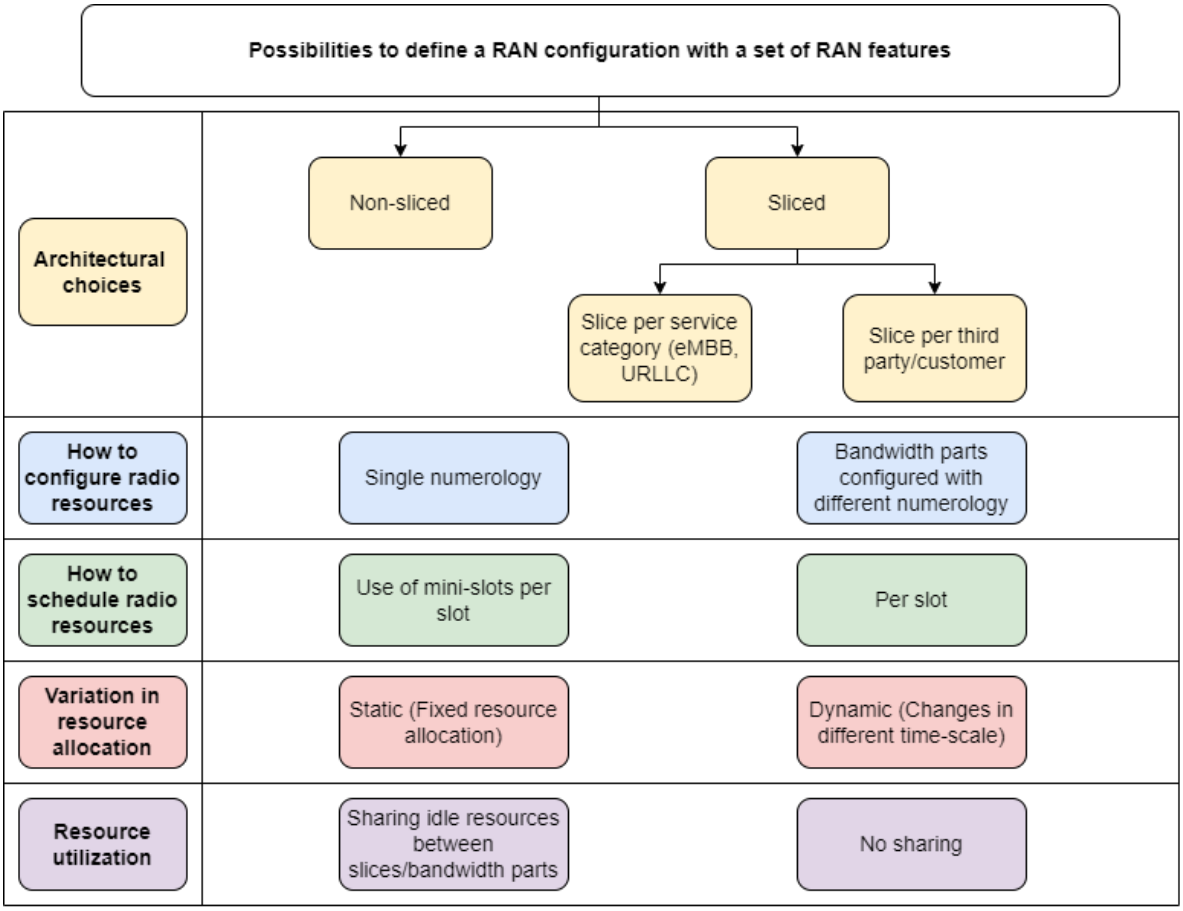

Figure 1.2: Choices of features to define a RAN configuration.

Each end-to-end slice comprises a RAN slice, a transport slice and a core network slice, as shown in Figure 1.3. In this study, the focus is on RAN slicing as the study focuses on optimizing the RAN for provisioning diverse service requirements in the same physical network.

Considering a network with heterogeneous service requirements, managing the radio resources is a challenging task in order to satisfy the diverse service requirements of the distinct applications. With RAN slicing, there would be multiple slices, with each slice dedicated to a particular service category or can be customized for a third party or a customer as mentioned in Figure 1.2. Specifically, a RAN slice for a customer has a Service Level Agreement (SLA) which is a contract between the service provider and the customer with distinct service requirements. A customer, for example can be a company from the automotive industry having an SLA with the network provider. In this study, emergency service group is considered as a third party customer with service requirements of incident-based applications, video surveillance and sensors measuring risk-sensitive measurements throughout the city.

Another concept is that of *flexible numerology*, introduced in 5G in which the Sub-Carrier Spacing (SCS) of the Orthogonal Frequency Division Multiplexing (OFDM) symbols can be flexibly set to 15×2^n kHz, where n is the numerology (integer value in range 0 to 4) that can be adjusted based on the service requirements. With an increase in numerology value, the slot duration reduces as shown in Figure 1.4. The flexible numerology and RAN slicing feature can be combined, and the resources can be configured with different numerologies, as mentioned as a choice of configuring the resources at RAN in Figure 1.2. The details about the benefits and/or losses of flexible numerol-

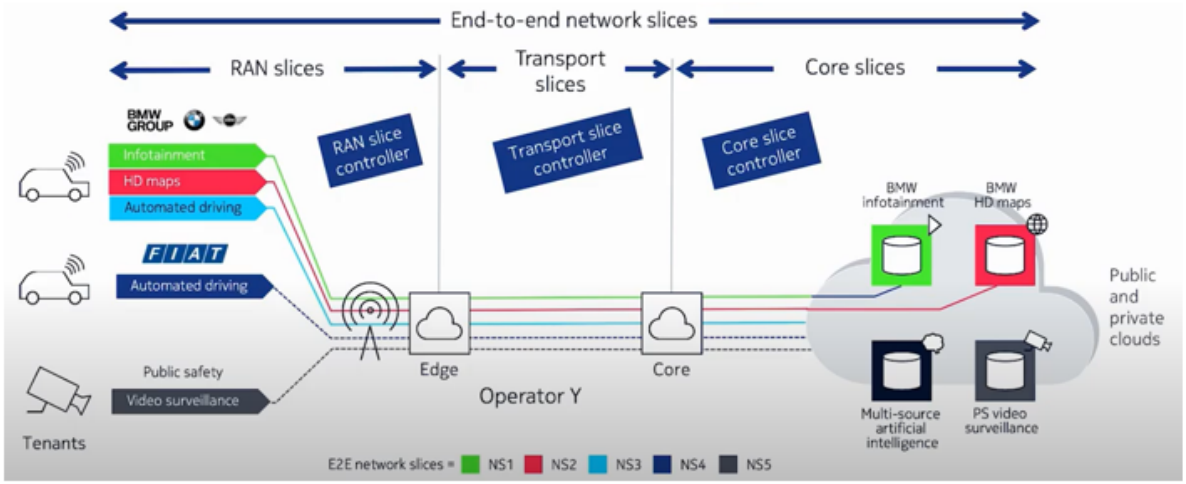

Figure 1.3: Network Slicing [12].

ogy and the combination of RAN slicing and flexible numerology feature in explained in the next chapter.

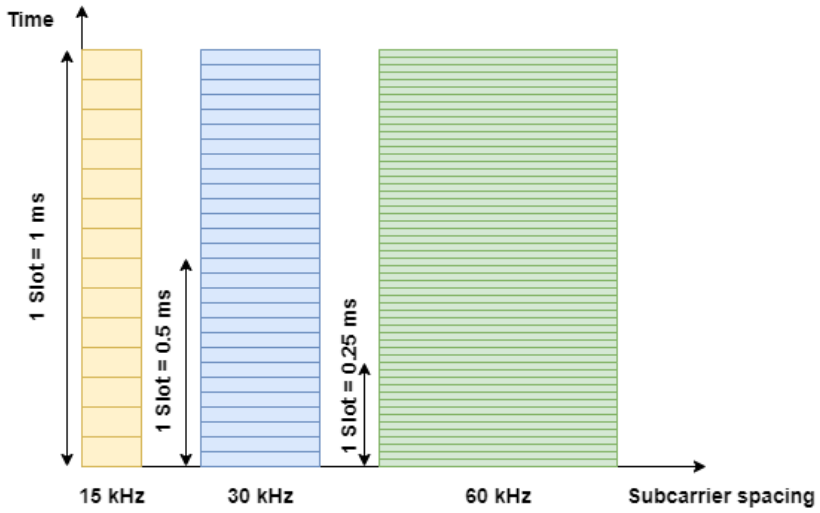

Figure 1.4: Flexible numerology [13].

Additionally, in 3GPP Rel-15 [14], the concept of *BWP* is introduced, where a BWP is part of the total carrier bandwidth. Each BWP can be configured with a different numerology. In other words, BWP is a way of configuring bandwidth with multiple numerologies in a non-sliced RAN or within a slice, as also mentioned as a possibility in Figure 1.2. The difference between the BWPs and slicing concept is explained in the next chapter.

In 5G New Radio (NR), a slot contains 14 OFDM symbols in the time domain, and similar to 4G RAN, transmissions are scheduled in slots. However, in 5G, mini-slots are introduced which can consist of two, four or seven OFDM symbols as shown in Figure 1.5 and they are, therefore, smaller than a regular slot. Transmissions in 5G can be scheduled using a regular slot or by using mini-slots as mentioned in Figure 1.2. Also, transmissions related to mini-slots can be scheduled at any time within a slot. The flexibility of using mini-slots for scheduling can be combined with the flexible numerology and/or slicing and/or BWPs features to optimize the RAN for QoS provisioning of distinct applications.

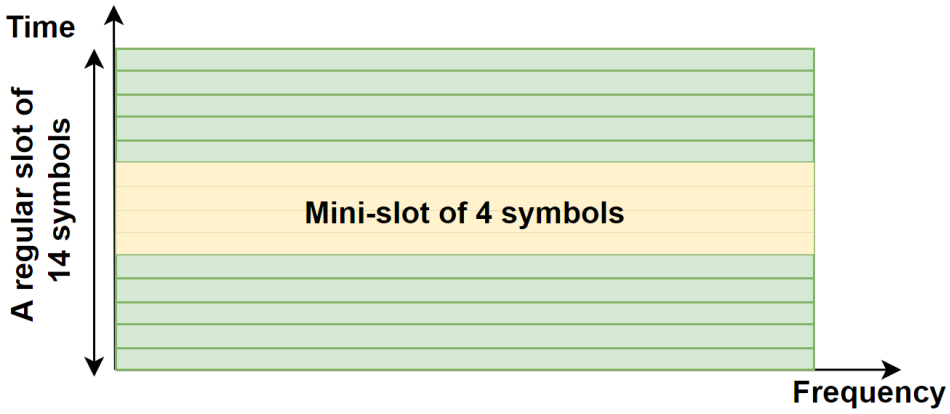

Figure 1.5: Mini-slot in a regular slot [13].

Considering that all slices share a common physical infrastructure, the radio resources are distributed among the slices, which can be done statically or dynamically, as mentioned in the Figure 1.2. If the traffic load offered to different slices structurally changes over time, then the adaptations should be made in the radio resource slice assignment. The timescale at which the slice resource assignment can be done could be high, such as hours or weeks. In principle, it is also possible to do the resource assignment at very fine timescales, such as milliseconds, but that approaches the timescale at which scheduling operates which is the same as differentiated scheduling and the distinction with slicing becomes meaningless [15].

Additionally, there is an option of sharing the idle resources between slices and/or BWPs. In sliced RAN and/or BWPs, the RAN resources are divided among the slices or BWPs which in case of no idle resource sharing reduces the multiplexing gains as the user per slice or BWP can only be served using the dedicated resources to its slice or BWP, leading to lower trunking efficiency in comparison with non-sliced RAN. Therefore, the idle resource sharing is considered as a possibility in this study.

1.4. Research Motivation and Challenges

The expectation from 5G is to support the eMBB, URLLC and mMTC service categories simultaneously on the same physical infrastructure. It is a very complex and challenging task to manage and optimize the RAN to support these services together due to their diverse QoS requirements. As previously mentioned, RAN slicing allows managing the network more easily, as each slice can be configured independently according to the slice-specific requirements, using the concept of flexible numerology, use of mini-slots for scheduling and allowing sharing of radio resources between slices. Even with non-sliced RAN, BWPs can be used to configure the RAN with multiple numerologies to support distinct applications. Hence, this thesis studies the potential of 5G RAN features for QoS provisioning of a multi-service smart city environment.

The challenge of the study is to identify and understand the impact of each 5G RAN feature and to overcome the drawbacks of one feature by combining with other features. The different features can be used individually or can be combined to improve the performance, which leads to many possible and meaningful RAN configurations. To derive the optimum among all possible combinations is the main focus of the study. Determining these large numbers of possible combinations for RAN configuration and then making the correct selection to analyze the impact of each feature individually or in combination with the other is a challenging task.

To assess the meaningful RAN configurations, an existing 5G system-level simulator is substantially upgraded. Beside the research challenges, there was a practical challenge in simulating these

types of complex scenarios and large number of RAN configurations. For statistically reliable results, multiple simulations are necessary for each considered RAN configuration, which brings the challenge of long simulation times and the need for a large amount of computational resources.

1.5. Related work

With the new emerging applications having different service requirements, 5G needs a new network architecture to fulfill these diverse requirements and manage the network. Elayoubi et al. [16] discuss the potential of RAN slicing in managing different service requirements and describe four options of configuring a RAN slice: (i) a slice for each service category (e.g. URLLC, eMBB), (ii) a slice for each set of technical requirements (e.g. 1 ms latency, 2 ms latency), (iii) a slice for each customer (e.g. automotive industry, offshore industry) and (iv) a slice for each customer per technical requirement (e.g. a customer such as automotive company with slice for safety messages for autonomous driving and slice for in-vehicle infotainment service).

Rost et al. [17] discuss a framing structure for sharing the time-frequency resources which is in line with the standardized BWPs feature and is called *tiling*. That is, the time-frequency resources are distributed in a tiling pattern and each tile is configured with some numerology. The RAN slices are then assigned resources from tiles with the appropriate numerology according to the service requirements of the slice. Based on the concepts of tiling, Sexton et al. [18] show four approaches for sharing radio resources between RAN slices, as shown in Figure 1.6. In two of the approaches, a contiguous sub-band is considered. At first, a pre-determined fixed adjacent sub-band region is used by each RAN slice. Each fixed region is separated by guard bands which are necessary to avoid Inter Numerology Interference (INI) which is caused due to non-orthogonality between different sub-carriers of different numerologies, as shown on the top left part of Figure 1.6. The second contiguous sub-band approach consists of a fixed and variable region, where the variable region comprises a ratio of two different numerologies separated by a guard band, as shown on the top right part of Figure 1.6. The variable region is shared between adjacent RAN slices. The other two approaches are based on the tiling concept. The one approach shows a sub-band tiling pattern in which the whole resource grid is split into sub-bands of a predetermined size, as shown in the lower left part of Figure 1.6. The other approach shows the frame tiling pattern in which the frames are configured with different numerologies and separated by guard bands, as shown in the lower right part of Figure 1.6. The four approaches are compared in terms of the number of required guard bands and their adaptability to the varying traffic and it is concluded that the variable contiguous sub-band approach is the most promising approach.

In the RAN, the appropriate amount of radio resources have to be allocated to each slice and/or BWP to support the SLAs, which is a complex task. Khatibi et al. [19] address this challenge by using Artificial Intelligence (AI) to learn and predict the traffic flow pattern of each slice and/or BWP. Specifically, they predict the traffic demands for the next scheduling slot and they then decide how many resources should be allocated to each slice and/or BWP.

Along with assigning resources to slices and/or BWPs, some studies also consider assigning resources directly to users. Li et al. [20] propose a two-level radio resource allocation framework, at the network-level and the gNodeB (gNB)-level, where the resource allocation at the two levels is done at different timescales. At the network-level, the radio resources are pre-allocated to the gNBs based on the service requirements. The resource allocation to gNBs varies over time, to adapt to the needs of time-varying traffic. At the gNB level, the pre-allocated resources of the gNBs are dynamically scheduled to the users at a mini-slot time-scale which is smaller than the time-scale at the network level. Also, the gNBs can share their idle radio resources with overloaded gNBs, which increases the overall resource utilization. Another way of allocating the radio resources to the RAN slices is proposed by Khodapanah et al. [21] who introduce a mapping layer and propose a slice-aware adaptation algorithm for the mapping layer. The mapping layer is a network entity that over-

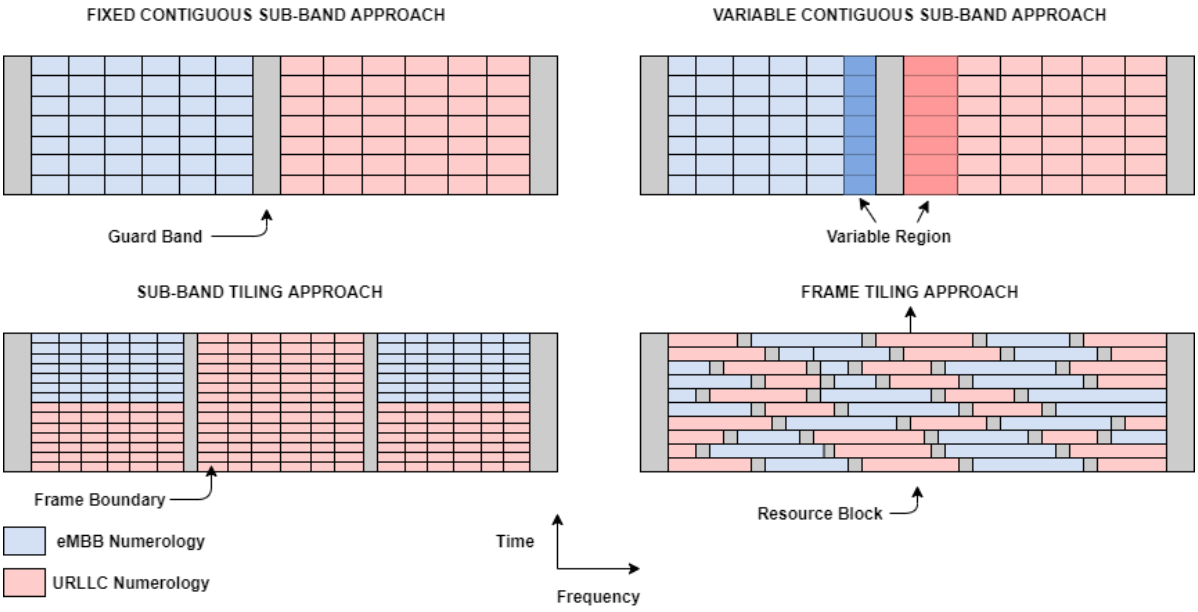

Figure 1.6: Four options for sharing time-frequency resource pool between two services employing different numerologies [18].

sees the network at the service area and manages the radio resource allocation between the slices to guarantee the targeted QoS requirements. To do so, the mapping layer keeps track of the slices' performance and tunes accordingly some weighting parameters related to packet scheduling. In particular, these weighting parameters are adjusted for every user and every slice as they are used for user prioritization during packet scheduling. The proposed adaptation algorithm uses a cost function for the target Key Performance Indicators (KPIs), as defined in the SLAs, and the experienced KPIs are tracked by the mapping layer. The corresponding cost values depict the deviation of the experienced KPIs from the target KPIs and therefore the adaptation algorithm aims to minimize the cost values.

Other studies address QoS provisioning without the use of RAN slicing. Pedersen et al. [22] present a punctured scheduler to multiplex URLLC and eMBB traffic on the DL shared channel. The proposed scheduler enables URLLC transmissions during ongoing or scheduled eMBB transmissions to minimize the latency of URLLC transmissions at the cost of throughput for eMBB transmissions. Specifically, the eMBB transmission with the lowest Modulation and Coding Scheme (MCS) is selected for puncturing to reduce the overall performance degradation of eMBB traffic. Zaidi et al. [13] provide a discussion about how mini-slots and multi-numerology can be used to serve traffic with different requirements (e.g. URLLC, eMBB) on the same carrier. In addition, it concludes that windowing and/or filtering signal processing techniques and insertion of guard bands can be used to reduce INI while multiplexing the RAN with multiple numerologies.

1.6. Research Objectives

Based on the application requirements considered in this study as well as the related work, discussed in the previous sections, the objective of this work is:

- Determine, configure and assess all the meangingful combinations and configurations of the 5G RAN features for provisioning a smart city multi-service scenario.
- Derive the conclusions on the merit of these different 5G RAN features.

1.7. Research Approach and Outline

The research approach and steps taken in achieving the objectives of this study as stated in Section 1.6 are:

- Identify the applications of smart city and derive their requirements.
- Perform a literature review and obtain a deep understanding of 5G technological features such as flexible numerology, the use of mini-slots, BWPs and RAN slicing for provisioning a multi-service environment.
- Derive different options of configuring the RAN by applying the above-mentioned 5G RAN features individually or by combining these features in a RAN to achieve the performance requirement of the diverse applications.
- Describe and model the investigation scenario such as network layout, propagation model, the traffic model for each considered application and the choices of the RAN configurations with different set of features.
- Substantially upgrade a pre-existing 5G system-level simulator from a single cell factory scenario to an urban macro-cellular environment scenario consisting of multiple cells, incorporate the UL transmission capability, add the considered applications of the reference scenario based on their traffic models and the options of mini-slot-based scheduling, BWPs, resource sharing between slices and/or BWPs in a sliced and non-sliced RAN, respectively.
- Run different simulations by configuring the RAN with different 5G RAN features as mentioned above, individually or in combination, and assess them based on the simulation results.
- Derive conclusions, explain the limitations and provide the future scope of this research.

2 5G NR Key Features

This chapter provides details on the 5G key enablers in achieving the diverse QoS requirements of the customer and the distinct applications. Section 2.1 gives the details on RAN slicing and describes the RAN slicing options. Section 2.2 explains the concept of flexible numerology and the need for guard bands to enable the multiplexing of different numerologies in the RAN. Section 2.3 describes how the bandwidth can be split into several BWPs. Further, Section 2.4 provides details on duplexing. In Section 2.5, the role of the packet scheduler and the options of mini-slot based scheduling are described. Section 2.6 provides details on the radio resource sharing between slices and BWPs. Finally, Section 2.7 provides a qualitative comparison between the possible RAN configurations using the 5G RAN features.

2.1. RAN Slicing

The 5G network is envisioned to be a multi-service and multi-tenant network that requires a flexible RAN to meet the diverse QoS requirements of the vertical sectors on a common physical infrastructure. This gives motivation for RAN slicing. With the concept of RAN slicing, the 5G spectrum can be divided into multiple parts of time-frequency resource grid which are assigned to different RAN slices and customized based on each RAN slice service requirement. The RAN can thus be configured with multiple RAN slices and a User Equipment (UE) can be assigned multiple RAN slices depending on the UE's service requirements. Third Generation Partnership Project (3GPP) has standardized four types of RAN slices, namely eMBB, URLLC, massive IoT (mIoT) and Vehicle-to-everything (V2X) slices [23] and also allows a network operator to flexibly define and configure additional RAN slices. The eMBB slice handles services which are throughput-oriented, URLLC slice handles services with low latency and high reliability requirements, mIoT slice handles the massive number of IoT devices requirement and the V2X slice handles services related to vehicles such as in-vehicle infotainment services, navigation services and communication of safety messages for autonomous driving.

In this study, based on the considered application requirements, the two options of RAN slicing are considered as explained below [16]:

- **A RAN slice per service category**: Considering the three generic service categories of 5G as mentioned in Chapter 1, namely eMBB, URLLC and mMTC, the RAN is configured with three RAN slices where each slice is dedicated to each service category, as shown in Figure 2.1.

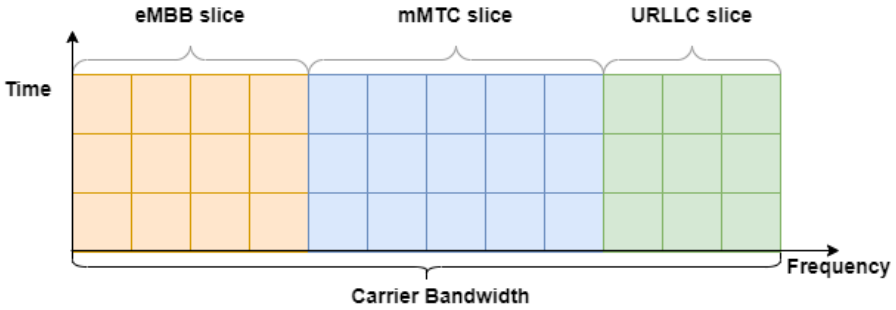

Figure 2.1: Example of RAN configured with three service category slices.

- **A RAN slice per service category and per customer**: Considering a customer like in this study,

the emergency service group, having an SLA agreement with specific QoS requirements, a separate RAN slice can be configured, dedicated to the customer, next to the three generic RAN slices, as shown in example Figure 2.2. This configuration helps the service provider to manage the customer's service demands together with the other generic service category requirements.

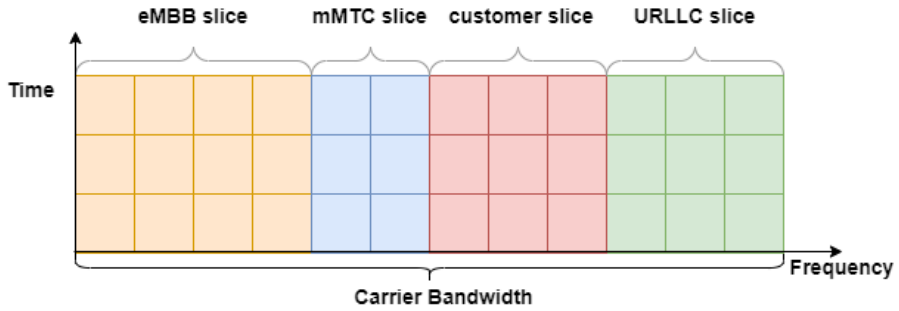

Figure 2.2: Example of RAN configured with customer slice together with three service category slices.

Each RAN slice should be optimally configured to meet the SLA. This flexibility of slicing the RAN enhances the manageability of QoS provisioning in a RAN, supporting applications of diverse service categories and distinct customers. A number of radio resources are assigned to each slice. Each slice has a packet scheduler that then assigns the slice's radio resources to the slice's users. Each slice can have a different packet scheduler based on a slice-specific requirement. However, because the traffic varies over time, the static resource assignment to each slice can lead to wastage or shortage of resources. This can be resolved by dynamically assigning the resources to each slice. In this study, the average traffic demand is constant; with variations only at finer timescale, which is handled by the scheduling mechanism and/or by idle resource sharing between RAN slices. Thus, dynamic slice resource assignment is out of focus for this study. The split of resources between RAN slices can lead to trunking losses in comparison with non-sliced RAN, in case of no idle resource sharing between slices, due to reduced multiplexing gains [24]. Thus, the idle resource sharing between slices is an important feature considered as an option in the study, which is explained in detail in Section 2.6.

2.2. Flexible Numerology

In 5G NR, the concept of flexible numerology is introduced. Unlike 4G, the SCS of OFDM symbols in 5G is not fixed to 15kHz. The SCS can flexibly be set to 15 x 2^{μ} kHz, where μ is the numerology value which is an integer in range from 0 to 4. The slot duration reduces with the factor of 2^{μ} with increase in numerology value, as shown in Figure 1.3 in the previous chapter.

In 5G, different frequency ranges are available. Specifically, the frequency band under 7.125 GHz is labeled as Frequency Range (FR) 1 and the frequency band above 24.25 GHz is labeled as FR2. The numerology value is limited based on the carrier frequency. Table 2.1 shows the relation between the numerology value, the SCS, the slot duration and the applicable frequency range as explained above.

The higher numerology values are more suitable for URLLC traffic as with a higher numerology, the slot duration is shorter. This helps in faster transmission of data. On the other hand, with a higher numerology, the total number of Physical Resource Blocks (PRBs) will be lower within a given bandwidth due to the wider SCS. Hence, the gains from frequency-selective channel-adaptive scheduling, explained further in Section 2.5, are reduced and the throughput is negatively affected. Therefore, a lower numerology is more suitable for traffic with a throughput requirement. This trade-off between the latency and throughput is considered while selecting the appropriate nu-

Table 2.1: 5G numerologies [25].

Numerology (μ)	**SCS (kHz)**	**Slot duration** (ms)	**Frequency range**
0	15	1	FR1
1	30	0.5	FR1
2	60	0.25	FR1 and FR2
3	120	0.125	FR2
4	240	0.0625	FR2

merology for each service category.

Different parts of the bandwidth or BWPs and different RAN slices can be configured with different numerologies to serve traffic with different requirements. A RAN slice can be configured with multiple numerology by splitting the assigned resource grid of the slice into multiple BWPs. However, multiplexing numerologies in the same carrier, introduces INI because the sub-carrier spacing of each numerology is different and thus sub-carriers from different numerologies are not orthogonal to each other as shown in Figure 2.3. Therefore, there will be an overlap of sub-carriers from different numerologies that interfere with each other. When only one numerology is used, all sub-carriers are orthogonal to each other and thus there is no interference. INI can be eliminated by using signal processing techniques, like windowing and filtering, and with a sufficiently large guard band between the sub-bands configured with different numerology [26]. Despite using windowing and filtering, a minimum guard band size is also necessary between different sub-carriers [26] as shown in Figure 2.4. The guard band sizes are shown in Table 2.2.

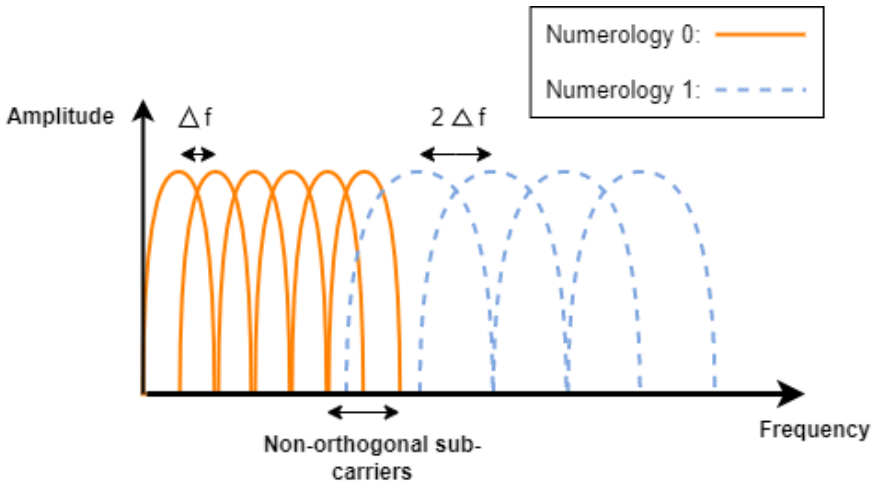

Figure 2.3: Inter-numerology interference between sub-carriers of different numerologies.

Table 2.2: Guard band size between sub-bands based on the adjacent sub-bands' numerology combination [26].

Adjacent numerologies	**Guard band size (kHz)**
0 and 1	150
1 and 2	300
0 and 2	120

Apart from the guard band between sub-bands with different numerologies, there are also edge guard bands, at the two edges of the frequency carrier, where each edge guard band size depends on the numerology configuring the relevant edge, as also shown in Figure 2.4. Table 2.3 shows the minimum edge guard band size based on the numerology and the carrier bandwidth [27]. For ex-

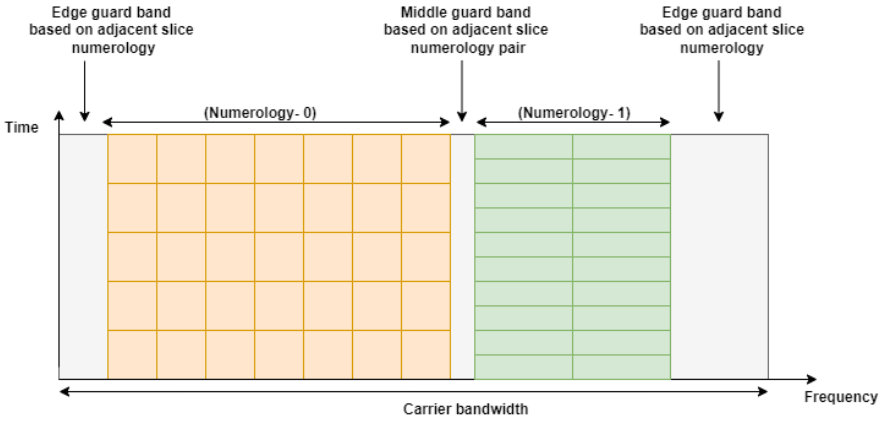

Figure 2.4: Example of RAN configured with two numerologies and the relevant guard bands.

ample, consider a carrier bandwidth of 10 MHz, which is configured with two numerologies (0 and 1). Two edge guard bands of 312.5 kHz and 665 kHz for each numerology, respectively, are needed and a middle guard band of 150 kHz to avoid INI which results in a loss of 11.27% of total carrier bandwidth.

Table 2.3: Minimum edge guard band based on the numerology and the carrier bandwidth [28].

Numerology value	**Minimum edge guard band size (kHz) for different carrier bandwidth**		
	10 MHz	**15 MHz**	**20 MHz**
0	312.5	382.5	452.5
1	665	645	805
2	1010	990	1330

2.3. Bandwidth Parts

As per 3GPP, the RAN or a RAN slice can be configured with multiple numerologies with the use of BWPs, A BWP refers to a sub-band of the total carrier bandwidth, which can be configured with a numerology [25]. This flexibility allows supporting multi-service traffic in the same carrier. For example, one BWP can be configured with higher numerology to serve the traffic with latency requirements and the other BWP can be configured with lower numerology to serve traffic with throughput requirement, as shown in Figure 2.5.

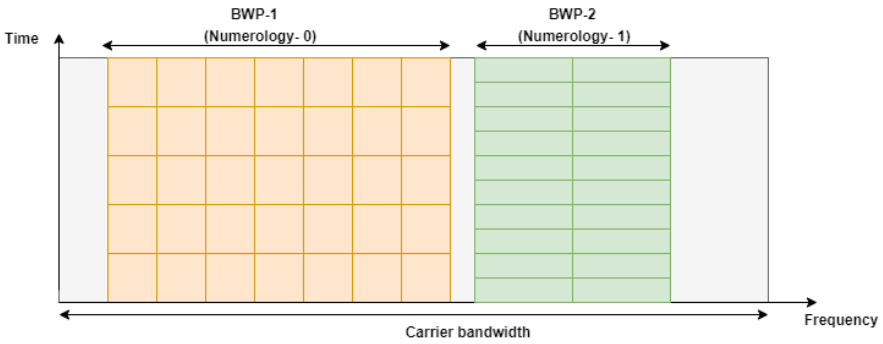

Figure 2.5: 5G-RAN configured with different numerology using BWPs.

A UE can be configured with a maximum of four BWPs, but only one BWP can be active, and

thus used for a transmission, at a given time [25]. A UE can switch to a different BWP at the cost of a switching delay which depends on the numerology of the initial BWP [29]. The BWP concept is similar to slicing with the key differences that unlike slicing, BWPs cannot have different packet scheduler per BWP and the Time Division Duplexing (TDD) configurations cannot differ per BWP as it is indicated and controlled by higher layers [30].

2.4. Duplexing

In wireless communication, duplexing is a process of achieving two-way communication over a communication channel, i.e. from Base Station (BS) to UE, the DL transmission over DL channel and from UE to BS, the UL transmission over UL channel. Duplexing can be distinguished into two types, namely Frequency Division Duplexing (FDD) and TDD, depending on whether the DL and UL channels are multiplexed in frequency or in time, respectively. With FDD, the UL and DL transmission is done using different frequency in the same time slot. While with TDD, the UL and DL transmission is done using the same frequency in different time slots. In this study, the TDD duplexing scheme is used, as it is standardized per frequency band that which duplexing scheme should be used, which in this study is taken as 3.5 GHz.

The slots used for the DL and UL transmissions are referred as DL and UL slots. The TDD configuration has a TDD frame size in number of slots which comprises a contiguous set of DL slots, then a flexibly configurable special slot (S), followed by a contiguous set of UL slots. Like the DL and the UL slot, a special slot (S) also contains 14 OFDM symbols. These 14 symbols in the special slot can be either DL or UL, as well as symbols that act as a Guard Period (GP) between the DL and UL symbols. The GP is the time interval required while switching from a DL to an UL transmission. This period is used to ensure that the already ongoing DL transmission is finished before the start of an UL transmission to avoid interference between the two transmissions at the BS. Contrarily, there is no need for a GP when transitioning from the UL to the DL channel because of the timing advance feature controlled by the BS and used by the user equipment [31]. Specifically, the BS indicates the time of an UL transmission, including the time caused by propagation delay depending on the corresponding distance between the BS and the UE. This time is controlled by the BS and is set in a way that the UL transmission finishes before the start of the DL transmission. For cell sizes up to 10.7 km, a GP of two symbols is considered sufficient and used in this study [31].

Unlike 4G, there are no predefined TDD patterns in a 5G radio frame[1]. With this flexibility in 5G, the optimal TDD configuration can be derived based on the traffic percentage of DL and UL traffic in the network. Additionally, in a sliced-network, each RAN slice can have a different TDD configuration. An example of how to derive a TDD configuration is given below.

Example: Consider a network with 75% of DL traffic and 25% of UL traffic. Assuming a TDD configuration with periodicity of five slots, there are a total of 14 x 5 = 70 OFDM symbols. With the reduction of two symbols used for GP, 68 symbols are available to be assigned to either the DL or UL channel. 75% of the 68 symbols i.e. 51 symbols are assigned to the DL channel and similarly, 25% of the 68 symbols i.e. 17 symbols are assigned to UL channels. As one slot comprises 14 symbols, this split translates to three DL slots, an S slot with nine DL symbols, two GP symbols and three UL symbols and one UL slot, as illustrated in Figure 2.6.

2.5. Scheduling

This section gives details about the scheduling mechanism at the RAN. The details about the packet scheduler and some commonly used packet schedulers are provided in Section 2.5.1. Additionally,

[1] In principle, this flexibility of selecting the TDD pattern in 5G exists, but the regulatory apply conditions via the spectrum licenses to use a given TDD configuration that consists of 8 DL and 2 UL slots. This condition is applied to avoid interference between different carriers used by different operators with different TDD frame configurations [32].

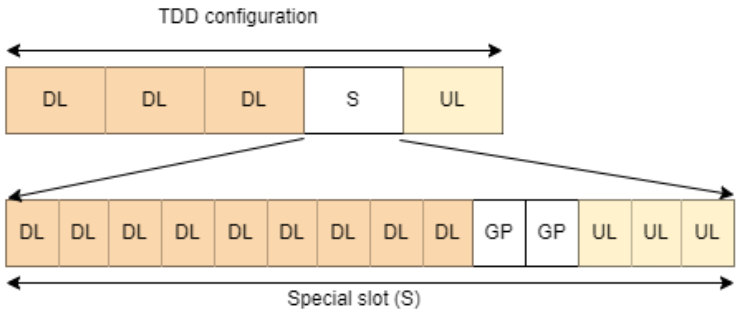

Figure 2.6: TDD example configuration.

mini-slot based scheduling can be used for faster data transmissions, as explained in detail in Section 2.5.2.

2.5.1. Packet Scheduler

The role of the packet scheduler is to assign the time-frequency radio resources to transmit the packets that arrived in the transmission buffer of the BS. The scheduler is developed and implemented by the network vendor and is not standardized. The scheduler in this study uses frequency-selective scheduling to leverage from the channel fading characteristics by allocating the best resources to the users in an optimal manner. The scheduler decides which packets to be served at which slot t and using which PRB f. A buffer is maintained of all the packets of each active user i for a given slot t. For each active user i, the current attainable bit rate is calculated based on the Channel Quality Indicator (CQI) feedback for each PRB f, in the given slot t. Then, a metric $M_{P,i}(t,f)$ is calculated for each head of the line packet of each active user i present in the buffer. The metric depends on the scheduling scheme of the chosen packet scheduler, labeled as P. For each PRB, at a given slot, the scheduler checks which head of the line packet in the buffer has the highest metric $M_{P,i}(t,f)$ and assigns that PRB to the corresponding active user i. This per-PRB level scheduling, at every scheduling slot, provides frequency diversity gains [33].

A packet scheduler also decides to not serve or drop a packet, depending on the conditions of the scheduling scheme. For example, for a latency-aware packet scheduler, the scheduler checks the latency budget of the packet. The latency budget is the maximum acceptable time frame in which the packet needs to be transmitted within the RAN, which is set based on the application requirements. The scheduler drops the packets which are not sent within the latency budget.

There are many packet schedulers that are designed to benefit a specific service category. A very high-level description of the commonly used packet schedulers are mentioned below:

- **Earliest Deadline First (EDF):** EDF scheduler is a latency-based scheduler that aims to deliver packets within the target packet latency constraint. It takes into account the remaining time within the latency budget while evaluating the M metric for the selection process. That active user is given PRBs which has the least amount of time remaining within the target latency budget.

- **Maximum Rate (MR):** MR scheduler is a throughput-oriented scheduler that aims to maximize the system throughput [33]. The evaluation of the M metric is done based on channel quality of the active users, as it takes into account the current attainable bit rate by every active user at the scheduling time. The PRBs are assigned to that active user which has the highest attainable bit rate at the scheduling time.

- **Proportional Fair (PF):** PF scheduler is also a throughput-oriented scheduler, as it also aims to maximize the system throughput depending on the channel quality of the active users. Un-

like MR scheduler, PF scheduler has more degree of fairness in terms of resource distribution among the active users. PF scheduler evaluates the M metric based on the experienced bit rate till the scheduling time for the selection process. The active user with low experienced bit rate is more likely to be assigned resources.

- **Modified-Largest Weighted Delay First (M-LWDF):** M-LWDF scheduler takes into account both latency constraints and channel-adaptive aspects while evaluating the M metric for the selection process. M-LWDF scheduler works on the same principle as PF scheduler for the throughput-oriented traffic. For latency-oriented traffic, M-LWDF scheduler evaluates the M metric as a weighted version of PF scheduler, where the weight factor depends on the latency aspects [24]. Therefore, M-LWDF is both latency and throughput-oriented scheduler.

The M-LWDF scheduler is selected and implemented in this study. The detailed explanation of the scheduler and the modelling aspects are provided in next chapter.

2.5.2. Mini-slots

A regular slot comprises 14 OFDM symbols, and a regular transmission is scheduled in one slot. A mini-slot is a smaller scheduling unit that can consist of two, four or seven OFDM symbols. With the use of mini-slots, a transmission can thus be scheduled faster than with a regular slot, as the actual transmission time is shorter. Therefore, mini-slot-based scheduling can be used for scheduling the URLLC traffic which has stringent latency requirements. The selected mini-slot duration is derived based on the DL or UL data size relating to the scheduled transmission and the number of assigned PRBs or vice-versa. In case, there is a choice that the transmission can be scheduled using a mini-slot of two symbols and four PRBs or four symbols and two PRBs, a scheduler may typically be designed to utilize fewer symbols (former option), as the goal is to have faster transmissions. In a slot, there can be multiple mini-slots, and thus multiple mini-slot transmissions can take place in one regular slot. A mini-slot transmission can be scheduled at any time during a regular slot to achieve low-latency because a packet transmission does not have to wait until the start of a slot, depending on different mini-slot based scheduling schemes, explained further in this section. Another benefit of using mini-slots is that it reduces interference, since the transmission may occupy only part of a slot rather than a whole slot.

The limitations of using mini-slots is that it should not span over two adjacent regular slots and that it should be aligned to the symbol boundaries of the regular slots [13]. For example, if a mini-slot of 7 OFDM symbols and another mini-slot of 4 OFDM symbols are used consecutively, there will be 3 remaining symbols in the regular slot. Therefore, only a mini-slot of 2 symbols can still be scheduled, leaving one symbol of the regular slot wasted as shown in Figure 2.7.

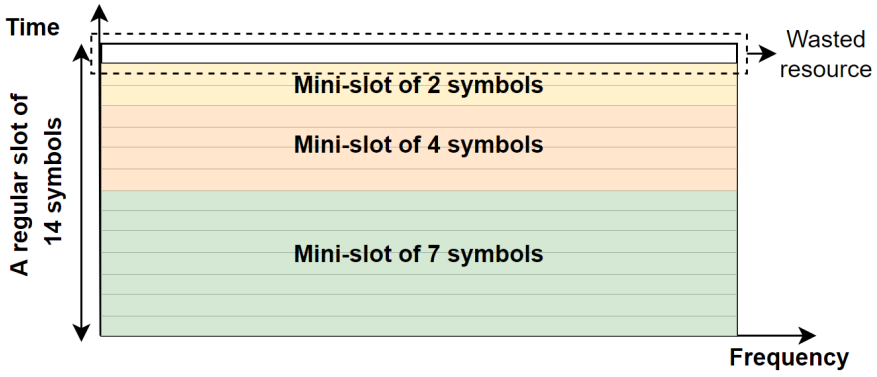

Figure 2.7: One example of mini-slot based scheduling.

There are different approaches when using mini-slots that are explained below [22][34]:

- **Basic mini-slot based scheduling:** The scheduler decides which packet from the transmission buffer should be served in the given scheduling slot, based on the *M* metric of the packet scheduler. The scheduler considers only the packets that are present in the buffer *at the start of a regular slot.* Then, the scheduler finds the minimum number of symbols needed for that transmission. To do that, the scheduler first checks whether the data can be transmitted with a 2 symbol mini-slot. If the transmission cannot be completed with 2 symbols, the scheduler checks whether the transmission can be completed with 4 symbols. Similarly, the process continues for 7 and 14 symbols (a regular slot). Once the minimum mini-slot duration is found, the transmission is scheduled. Then the next packet is selected, the cycle continues until there are no packets in the buffer or there are no more resources left for scheduling a transmission.

- **Non-pre-emptive mini-slot based scheduling:** In non-pre-emptive mini-slot based scheduling, there are two distinct types of scheduling moments: at the start of a regular slot and during a regular slot.

  At the start of a regular slot: The scheduler first schedules the packets based on the *M* metric of the scheduling scheme of the packet scheduler, similar to basic mini-slot based scheduling, considering the packets that are present in the buffer at the start of a regular slot.

  During a regular slot: After scheduling the packets at the start of a regular slot, as time progresses during the slot, the scheduler keeps checking the buffer for a new URLLC packet arrival. Specifically, the URLLC packets are prioritized as these packets have a latency requirement and should be scheduled as soon as possible. If a URLLC packet arrives during the already scheduled regular slot, the scheduler will check whether there are any unused resources remaining in the current slot, that were unassigned at the previous scheduling moment. If there are, the urgent URLLC data are scheduled for transmission in however many of those idle resources needed using a mini-slot. Consider the example shown in Figure 2.8 where e, u and m refers to eMBB, URLLC and mMTC packets, respectively, a new URLLC packet (u2) arrives at the buffer at 1.5 ms. After the scheduling round at the start of the regular slot i.e. at 1 ms, there were unused resources, represented with the white color in the figure, which can be now assigned to the newly arrived URLLC packet, using a mini-slot of appropriate length, which in this example is equal to four symbols. If there were no unused resources, the URLLC packet would have to wait until the next scheduling round, at the start of the next regular slot.

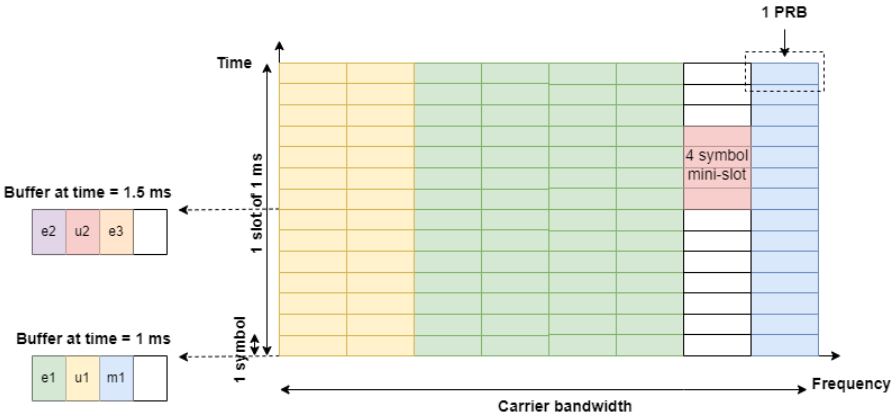

Figure 2.8: Example of non-pre-emptive mini-slot based scheduling.

- **Pre-emptive mini-slot based scheduling:** In pre-emptive mini-slot based scheduling, similar to non-pre-emptive and basic mini-slot based scheduling, there is again a distinction between two scheduling moments: at the start of a regular slot and during a regular slot.

  At the start of a regular slot: This scheduling moment is exactly the same as for the case of basic and non-pre-emptive mini-slot based scheduling.

  During a regular slot: After scheduling the packets at the start of a regular slot, the scheduler keeps checking the buffer for a new URLLC packet arrival. If a URLLC packet arrives during the already scheduled regular slot, the scheduler first checks for unused resources to assign it to the newly arrived URLLC packet. If there are no unused resources, the scheduler pre-empts an ongoing eMBB or mMTC transmission and schedules the urgent URLLC data using a mini-slot. The scheduler pre-empts those frequency resources for which the M-metric of the newly arrived URLLC packet is maximised, as it ensures the lowest amount of used resources and, consequently, the lowest degree of preemption of ongoing transmissions. The mini-slot duration is decided in the same manner as described in the above case. For example, consider that a URLLC packet arrives at time = 1.5 ms as shown in Figure 2.9 labeled as u2. Because there are no unused resources left from the scheduling round at time = 1 ms, one of the ongoing eMBB or mMTC transmissions needs to be pre-empted. Assume that the resources used for the eMBB packet number 2 denoted as e2 in Figure 2.9 has the highest M metric for the new URLLC packet. Then, the scheduler will pre-empt the e2 packet and assign the required amount of resources to the URLLC packet using a mini-slot, which in this example is equal to four symbols, as marked with the red color in Figure 2.9. The remaining resources of the pre-empted transmission are wasted and the pre-empted packets needs to be re-transmitted.

  An alternative option of pre-emption would be to pre-empt that ongoing eMBB or mMTC transmission which consumes the least number of PRBs for their own transmission. This way there will be the least possible loss of resources due to preemption. The former option is selected is this study.

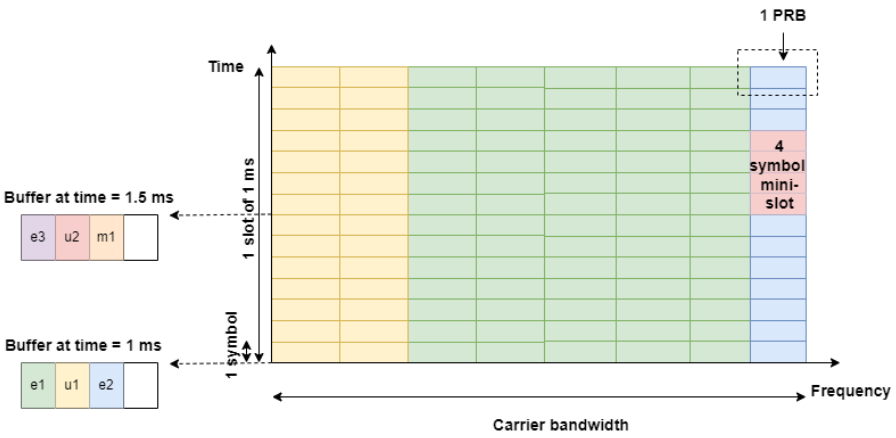

Figure 2.9: Example of pre-emptive mini-slot based scheduling.

The basic mini-slot based scheduling is a more efficient way of utilizing the radio resources in comparison with regular slot based scheduling. The non-pre-emptive and pre-emptive mini-slot based scheduling help in achieving the URLLC traffic latency requirements. The key difference between the basic and non-pre-emptive mini-slot based scheduling is that with non-pre-emptive mini-slot based scheduling, the URLLC packets arriving during a regular slot has a chance to get scheduled earlier than the start of the next regular slot without any loss of ongoing transmissions. This advantage increases the probability of transmitting the URLLC packets within the latency bud-

get in comparison with basic mini-slot based scheduling. With pre-emptive scheduling, the URLLC packet arrived during a regular slot will definitely be scheduled at the same time of arrival if there is an ongoing eMBB or mMTC transmission. There is a drawback with pre-emptive scheduling that the eMBB and mMTC traffic is negatively affected due to preemption.

2.6. Inter-slice/BWP Radio Resource Sharing

The 5G-RAN can be configured with multiple RAN slices and/or BWPs and assigning a fixed amount of resources to each slice and/or BWP can in theory lead to a lower spectral efficiency, as also mentioned in Section 2.1 and demonstrated in [24]. Due to variability of traffic, not all the radio resources in a RAN slice and/or BWPs are used. Such idle resources of each slice and/or BWP can be shared with other slices and/or BWPs in need of resources at every time slot. This flexibility increase the possibility of achieving the slice or BWP service requirement, as well as helps in increasing the spectral efficiency by not wasting the resources. The description below for sharing resources between RAN slices and is the same for sharing resources between BWPs.

From the implementation point of view, in every scheduling time slot, the packets of each RAN slice are scheduled. If a RAN slice have unused resources, which can be used by another RAN slice which still have traffic in buffer to serve and thus would benefit with having additional resources. The packets of the RAN slice in need of more resources will then be scheduled based on the configured numerology of the RAN slice which has unused resources. As different RAN slices can be configured with different numerology and can have different TDD configurations, there are a number of cases to be considered while sharing the idle resources between RAN slices. These cases are described below:

- Resource sharing when a RAN slice configured with higher numerology utilizes resources of a RAN slice configured with lower numerology. Assume the same transmission direction in both RAN slices.

  For example: Consider two RAN slices: RAN slice 1 is configured with numerology 0, and thus its scheduling slot duration is 1 ms, and RAN slice 2 is configured with numerology 1, and thus its scheduling slot duration is 0.5 ms as shown in Figure 2.10. Therefore, slice 2 has twice as many scheduling opportunities as slice 1, in a given time interval.

  After the first scheduling moment of slice 1, there are unused resources in slice 1, as represented with white resource blocks in the figure. If there are packets in a buffer corresponding to a UE served in slice 2 which need resources, the scheduler can use the unused resources of slice 1. During an ongoing first scheduled transmission of slice 1, slice 2 has a second scheduling moment. If there are still packets in the buffer of slice 2 after the second scheduling moment at slice 2 (for example, packet P4 is still in the buffer of slice 2 as shown in figure), the scheduler cannot schedule packet P4 on the unused resources of slice 1. This is because the packet P4 arrived in the buffer of slice 2 at time = 0.5 ms and scheduler at slice 1 do not have a scheduling moment at that time. Thus, under this sharing option, the sharing of resources is only possible, when the time slots are aligned between the two slices.

  With the use of mini-slot based scheduling, this drawback can be mitigated. The scheduler at slice 1 can schedule at any moment in time using mini-slots, so now the scheduler can assign the unused resources of slice 1 to slice 2 packets at every scheduling moment. For example, in the above case, packet P4 of slice 2 can be assigned unused resources of slice 1 using the last seven symbols of the regular slot of slice 1. The resources that packet P4 will get will be configured with the numerology value of slice 1.

- Resource sharing when a RAN slice configured with lower numerology utilizes resources of a RAN slice configured with higher numerology. Assume the same transmission direction in

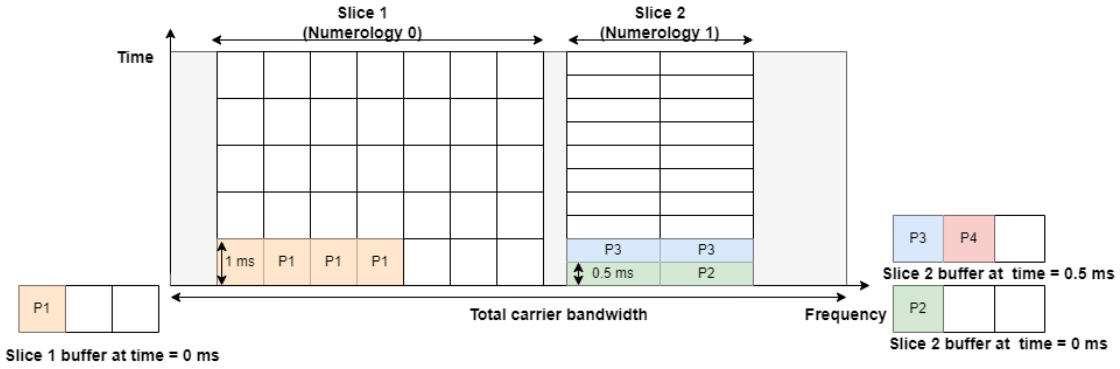

Figure 2.10: Example of the utilization of idle resources of a slice with a lower numerology by a slice with a higher numerology.

both RAN slices.

For example: Consider the same numerology configuration for both RAN slices as in case 1 as shown in Figure 2.11. In this case, after the first scheduling moment at both slice 1 and 2, there is still a packet P2 in the buffer of slice 1. If there were unused resources in slice 2, the scheduler will assign resources of slice 2 to the packet P2 of slice 1 by using the configured numerology of slice 2. This is possible only when the scheduling moment for both the slices are aligned in time. But in this case, there are no free resources after the first scheduling moment in slice 2 as shown in Figure 2.11. The slice 2 will have a second scheduling moment while slice 1 has an ongoing scheduled first transmission. There are resources left in slice 2 after the second scheduling moment as shown by white color resource blocks in the figure, those resources cannot be used to serve slice 1 packet P2 because the scheduler of slice 1 cannot indicate to the scheduler of slice 2 that there is a packet in the buffer of slice 1 because the scheduler of slice 1 does not have a scheduling moment at time = 0.5 ms.

With the use of mini-slot based scheduling, this drawback can be mitigated. The scheduler at slice 1 can schedule at every moment in time using mini-slot based scheduling. Therefore, the scheduler of slice 1 can indicate if there is a packet in the buffer of slice 1 at every scheduling moment. For the above case, the scheduler of slice 1 can now indicate the presence of packet P2 to the scheduler of slice 2. Then the scheduler can assign the unused resources of slice 2 to packet P2 of slice 1 using the configured numerology of slice 2 at the second scheduling moment of slice 2.

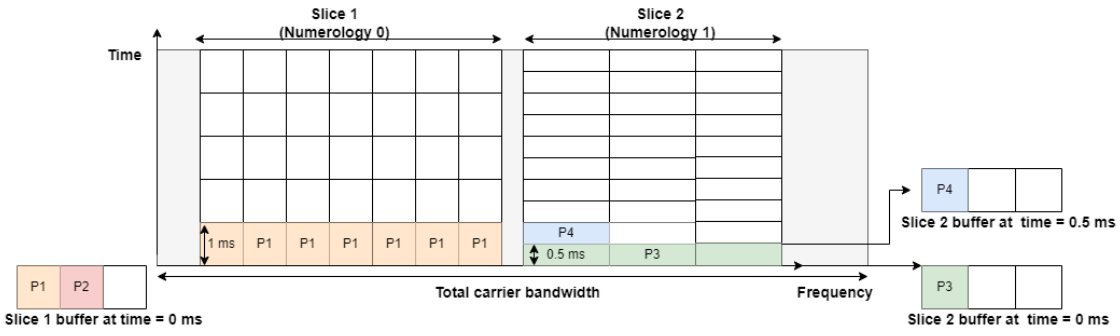

Figure 2.11: Example of the utilization of idle resources of a slice with a higher numerology by a slice with a lower numerology.

- Resource sharing when the transmission directions differ in both slices at the scheduling time.

For example: Consider RAN slice 1 and 2 that have different TDD configurations and they both have DL and UL packet transmissions. Now consider that the first scheduling slots of RAN slice 1 and 2 are a DL and an UL slot, respectively. If RAN slice 1 has unused resources and RAN slice 2 has some DL data in a buffer, the scheduler can assign the unused resources of RAN slice 1 to the DL packets in the buffer of RAN slice 2, because the slot of RAN slice 1 is dedicated to DL transmissions. Only the users of slice 2 which do not have a scheduled UL packets are assigned the resources of RAN slice 1 as the same user can either have a DL or UL transmission at a given time. The vice-versa is also true in which case there will be an UL data available in slice 1.

2.7. RAN Configurations

Each 5G RAN feature mentioned in the above sections can be exploited in achieving the different service requirements. The RAN may be configured with one or with a number of these features.

Taking into account the concept of flexible numerology, the RAN can be configured with one suitable numerology or it can be divided into multiple BWPs with different numerologies, based on the service requirements. Similarly, the RAN can be split into multiple RAN slices with all slices having the same numerology to reduce guard bands or with a different numerology per slice. Also, a RAN slice can be further divided into BWPs if within a RAN slice there are different service requirements, e.g. for a slice concerning a customer. Additionally, with RAN slicing, each slice can have a different TDD configuration and packet scheduler to achieve its target performance requirements.

The packet scheduling can be done using a regular slots or mini-slots can be used to efficiently use the radio resources. Additionally, non-preemptive or preemptive mini-slot based scheduling can be performed to prioritize the latency-constrained traffic. Mini-slot based scheduling can be done in combination with a suitable numerology to further improve the performance. Furthermore, to mitigate the drawback of using higher numerology for throughput-oriented traffic, the RAN can be configured with lower numerology and mini-slots can be used for traffic with strict latency requirements.

The drawback with splitting the resources between BWPs and/or slices can be reduced by sharing the idle resources between BWPs and/or slices.

The different possible RAN configurations are derived using these 5G RAN features and are presented in Table 2.4. Also, the advantages and disadvantage of each RAN configuration are explained in each table.

Table 2.4: Different combination of 5G RAN features to configure the RAN.

Combinations	Advantages	Disadvantages
Fixed numerology in a non-sliced RAN or for all RAN slices	- No use of guard bands.	- There is no single numerology suitable for different service requirements.
Different numerology between BWPs or RAN slices	- The BWPs or RAN slices can be configured with different numerology appropriate for each service category.	- Use of guard band to avoid INI between BWPs or RAN slices. - Selection of suitable numerology for a slice with mixed traffic requirements. - One packet scheduler for a slice with mixed requirements.
Different numerology between RAN slices and within a RAN slice	- The RAN slices can be configured with different numerology appropriate for each service requirement. - The RAN slice can further be configured with different numerologies to handle mix of traffic requirements.	- Use of guard bands between slices and within a slice to avoid INI.
Mini-slot based scheduling with fixed numerology in a non-sliced RAN or for all RAN slices	- Mini-slots can be used with fixed numerology to achieve strict latency requirement of URLLC traffic. - No need of guard bands.	- Performance degradation for ongoing transmission in case of pre-emptive scheduling. - In case of non-pre-emptive scheduling, delay in transmitting URLLC packets due to ongoing transmission.
Mini-slot based scheduling with different numerology between BWPs or RAN slices	- As URLLC BWP has different latency requirements, use of mini-slots can help in achieving the stricter latency requirement. - As a slice can have mixed traffic, use of minislots can help in achieving the strict latency requirement in that slice. - Avoid use of guard band within a slice.	- Use of guard band between BWPs or RAN slices to avoid INI. - Performance degradation for ongoing transmission in case of pre-emptive scheduling. - In case of non-pre-emptive scheduling, delay in transmitting the stricter latency URLLC data.
Mini-slot based scheduling with different numerology between slices and within a slice	- For the slice having different latency requirements, mini-slots can be used in achieving stricter latency requirement within that slice.	- Use of guard band between slices to avoid INI. - Performance degradation for ongoing transmission in case of pre-emptive scheduling. - In case of non-pre-emptive scheduling, delay in transmitting the stricter latency URLLC data.
Common advantage in all RAN slice combinations: Can assign different suitable packet scheduler and TDD configuration per slice depending on service requirement of each slice.		
Common disadvantage in all combinations: Reduced multiplexing gains due to splitting of radio resources between BWPs or RAN slices which can be improved by inter-slice and inter-BWP sharing of idle resources.		

3 Simulation Modelling

This chapter discusses the simulation modeling aspects, assumptions, and simulation flow for the implementation of the considered RAN features and the chosen investigation scenario. In Section 3.1, the details about the network topology are presented. In Section 3.2, the details about the propagation environment are presented, followed by the traffic models used, in Section 3.3. Based on the traffic models, Section 3.4 provides details about the network bandwidth estimation and calculation. The details of the different considered scenarios and their configurations are provided in Section 3.5. Section 3.6 explains the DL and UL transmission procedures. Section 3.7 concludes the chapter with an outline of the simulation flow.

3.1. Network Topology

The thesis focuses on an urban macro-cellular test environment as specified by 3GPP [35]. The network layout consists of 19 macro sites with 3 sectors per site. A total of 57 cells arranged in a hexagonal layout are available in the network, as illustrated in Figure 3.1. The Inter-Site Distance (ISD) is 500 m as mentioned in 3GPP guidelines [35].

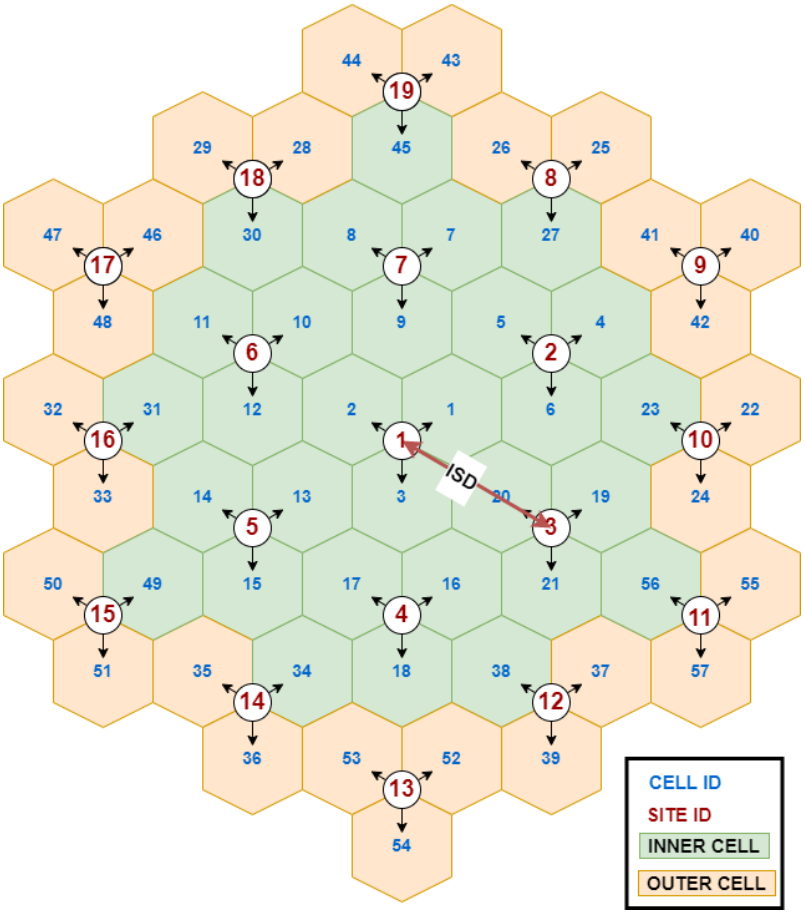

Figure 3.1: Network Layout.

The BSs are placed at the site locations with an antenna height of 25 m. Each BS serves three sectors and each sector has a directional antenna with an electrical downtilt of 10 degrees, based on [36], and a maximum antenna gain of 17 dBi. Figure 3.2 visualizes the antenna diagram obtained by QuaDRiGa model. The maximum transmit power for each sector of the BSs is 49 dBm and they have a noise figure of 3 dB, as also shown in Table 3.1.

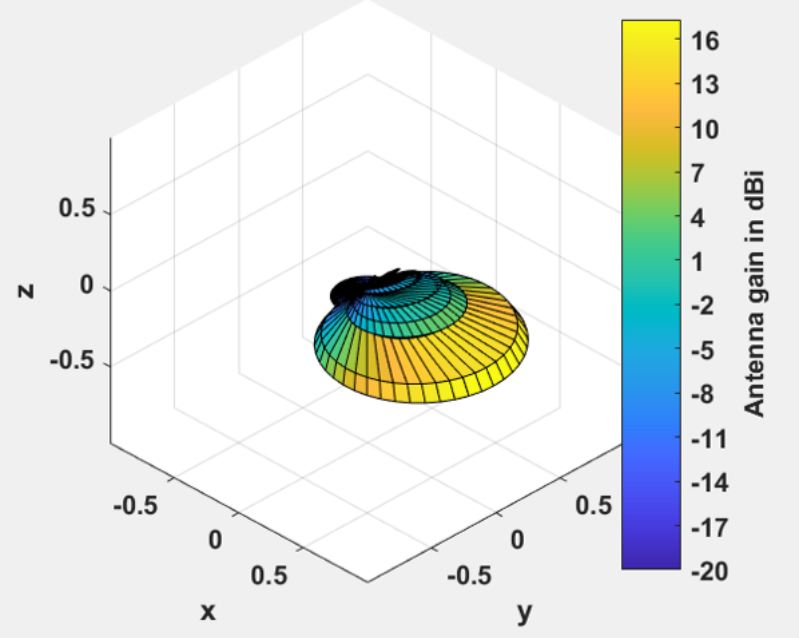

Figure 3.2: Antenna array pattern.

Table 3.1: Parameters for the base stations and the devices [35].

Parameters	Base station	Devices
Height	25 m	1.5 m
Maximum antenna gain	17 dBi	0 dBi
Transmit power	49 dBm	23 dBm
Noise figure	3 dB	9 dB

There are three types of devices in the network. Devices handling eMBB traffic, mMTC traffic and URLLC traffic. Each type of device is uniformly distributed in the network and they are located at a height of 1.5 m. As per 3GPP guidelines, the devices are located at a minimum distance of 35 m from the BS [35]. The devices have an omni-directional antenna with antenna gain of 0 dBi. The maximum transmit power of each device is 23 dBm and each device has a receiver noise figure of 9 dB, as also shown in Table 3.1.

3.2. Propagation Environment

The propagation environment is based on a densely built smart city scenario which consists of multiple obstacles for signal transmission such as buildings and trees. The obstacles can scatter or completely block the transmitted signal. These effects are taken into consideration in the multi-path fading and shadowing values. Following the guidelines by 3GPP in [35], the Path Loss (PL) is evaluated using the following equations given below. This study considers the Non-Line Of Sight (NLOS) links for the evaluation, to test the scenarios in harsh propagation environment.

The path of propagation which is unobscured is called Line Of Sight (LOS) path and which is obscured by an obstacle is called NLOS path. The transmission path can be divided into LOS and NLOS paths, so the path loss (PL_{NLOS}) of a transmission is defined as the maximum of the LOS and NLOS paths which is given by:

$$PL_{NLOS} = \max(PL_{LOS}, PL'_{NLOS}) (\text{for } 10\text{m} \leq d_{2D} \leq 5\text{km}) \tag{3.1}$$

where, the path loss (PL_{LOS}) for the LOS path is:

$$\begin{aligned} PL_{LOS} &= \begin{cases} PL_1 & 10\text{m} \leq d_{2D} \leq d_{BP} \\ PL_2 & d_{BP} \leq d_{2D} \leq 5\text{km} \end{cases} \\ PL_1 &= 28.0 + 22\log_{10}(d_{3D}) + 20\log_{10}(f) \\ PL_2 &= 28.0 + 40\log_{10}(d_{3D}) + 20\log_{10}(f) - 9\log_{10}((d_{BP}^2 + (h_{BS} - h_{UT})^2) \end{aligned} \tag{3.2}$$

where $f = 3.5$ GHz is the carrier frequency, d_{2D} and d_{3D} are the two and three-dimensional distance between the BS and the device and h_{UT} and h_{BS} are the device and BS heights, as shown in Figure 3.3. Further, d_{BP} is the breaking point distance which is defined by:

$$d_{BP} = 4h'_{BS}h'_{UT}f/c \tag{3.3}$$

where, the parameter c = 3 $\times 10^8$ m/s is the propagation velocity and h'_{BS} and h'_{UT} are the effective antenna heights at the base station and the device respectively, as also shown in Figure 3.3 and they are computed as follows:

$$\begin{aligned} h'_{BS} &= h_{BS} - h_E \\ h'_{UT} &= h_{UT} - h_E \end{aligned} \tag{3.4}$$

where h_E =1 m is the effective environment height for urban macro-cellular environments [35]. This effective environment height is dependent on environmental parameters such as vegetation depth, street width and location of the street. The breaking point distance obtained with the above-defined value of the parameters is 560 m.

The path loss (PL'_{NLOS}) for the NLOS path is:

$$PL'_{NLOS} = 13.54 + 39.08\log_{10}(d_{3D}) + 20\log_{10}(f) - 0.6(h_{UT} - 1.5) \tag{3.5}$$

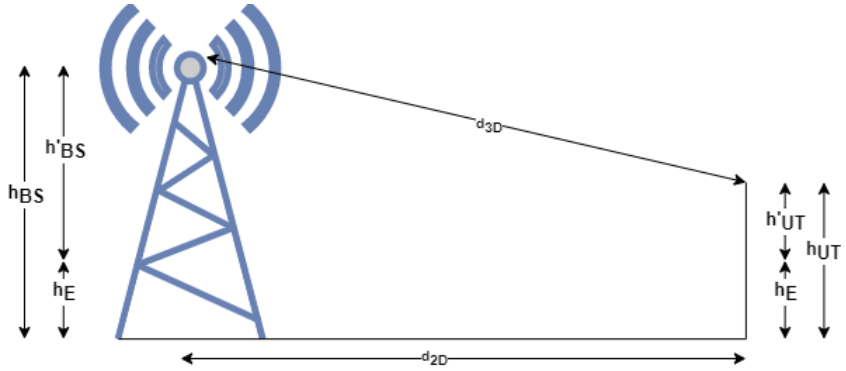

Figure 3.3: Parameters of BS and device as defined in the network [35].

Each radio link between a cell and a device is characterized by shadowing and multipath fading because of scattering of signal due to buildings in the smart city environment. The channel coefficients capturing the effects of multipath fading and shadowing for each radio link are generated with the QuadRiGa 3GPP Urban Macro-cell model [37]. The multipath fading follows a Rayleigh fading distribution and lognormal shadow fading is assumed with a standard deviation σ_{SF} = 6 dB [35].

Multiple multipath traces are pre-generated using the Quadriga model, independent of location of the cells and the devices. For each cell-device link, one of the pre-generated trace is randomly selected with a randomly selected starting slot of the trace. When the end of the trace is reached, the trace is wrap-around again.

3.3. Traffic Models

The applications considered in this study and their requirements are mentioned in Section 1.2. There are two general type of sessions in the network, as enlisted below:

- **Persistent sessions:** There are a N_{app} number of sessions of this type which are always present and are uniformly distributed in the network area. The app in notation refers to the specific application.
- **Non-persistent sessions:** The sessions of this type arrive in the network area following a spatially uniform Poisson process, with an arrival rate of λ_{app} sessions/second. These sessions are divided into two subtypes based on their session time period, as enlisted below:
  - Non-persistent session with fixed session time period of x_{app} seconds.
  - Non-persistent sessions that stay in the network until the file is transmitted completely.

In general, there are three types of processes specifying the arrival of packets within a session, as enlisted below:

- **Bulk arrival:** A file with a deterministic size of y_{app} MB for download or upload.
- **Periodic arrival:** The inter-arrival time between successive packets is fixed to T_{app} seconds. For each such packet flow, the arrival time of the first packet of the flow is chosen randomly in [0, T_{app}] seconds. This is the case with persistent sessions. For non-persistent sessions with periodic traffic, the arrival time of the first packet is the same as the arrival time of the session.
- **Poisson arrival:** The packets arrive following a Poisson process with an arrival rate of λ_{app} packets/second.

The traffic model details per application type are given below and the categorization of the applications in accordance with the above-mentioned session arrival and packet arrival processes are summarized in Table 3.2.

- **Broadband access everywhere:** The 3GPP File Transfer Protocol (FTP) Model 1 is considered as the most appropriate traffic model [1]. The sessions are non-persistent, having an arrival rate of λ_{BB-DL} = 100 sessions/second for DL and λ_{BB-UL} = 40 sessions/second for UL. The download and upload are done by the users at homes, offices and universities. Each session has a fixed download or upload traffic of 2 MB and thus a download/upload ratio of *2.5 : 1* is considered [38].
- **VR:** The non-persistent sessions for the VR gaming application have an arrival rate of λ_{VR} = 40 sessions/second. Each session is assumed to be 5 seconds long. In practice, VR sessions are likely to last much longer, but for simulation purposes, a choice is made to have many short VR sessions rather than a few long ones, to ensure sufficiently representative and statistically reliable simulation results. The VR gaming sessions are played around the city in homes and sport centers. The packets containing VR video content are generated according to a Poisson process with an arrival rate of λ_{VR} = 200 packets/second [39] and the packets containing VR motion feedback data arrive periodically with an interval of T_{VR} = 0.004 s [40].

- **Video surveillance:** There are 676 persistent CCTV cameras (200 cameras/km^2) around the city streets, which periodically send video chunks of 4.5 kB to their respective control centers with an interval of T_{VS} = 0.036 s [1][6].
- **Sensors for non-risk-sensitive measurements:** A total of 472920 such sensors (140000 sensors/km^2 [6]) are persistent in the network around the city streets and modelled to periodically transmit data of 200 bytes with an interval of T_{S-NRS} = 60 s [1][6].
- **Sensors for risk-sensitive measurements:** A total of 202680 such sensors (60000 sensors/km^2 [6]) are persistent in the network around the city buildings and houses and modelled to transmit data periodically of 200 bytes with an interval of T_{S-RS} = 60 s [1][6].
- **Sensors for emergency messages:** As mentioned in Section 1.2, a scenario with an incident is also considered. An incident area of 98 m^2 is considered, which is a typical size area of a commercial building [41]. During the incident, a total of 25 persistent sensors are assumed to be activated at the incident area to send emergency messages. The packets containing emergency data follow a Poisson process with a packet size of 32 bytes and an arrival rate of λ_S = 0.03 packets/second [1].
- **Body camera:** At the incident area, a total of two emergency workers are assumed to use body cameras during the incident for surveillance purposes. The cameras send video chunks of 4.5 kB periodically with an interval of T_{BC} = 0.036 s [1].
- **AR:** For rescuing people in the building, two emergency workers are assumed to use AR headsets during the incident, which periodically transmit motion feedback messages with size of 500 bytes with an interval of T_{AR} = 0.004 s [40]. The packets containing video updates of 1250 bytes arrive at the headset screen following a Poisson process with an arrival rate of λ_{AR} = 200 packets/second [39].

As mentioned before, in this study, we are also interested in analyzing the impact of the incident in the network. In order to do so, two distinct scenarios are defined and simulated separately, viz. one with and one without an incident. In the former case, the incident last for the whole simulation period. The incident area is randomly placed within an incident-cell which is randomly selected among the 57 cells in the network, as shown in Figure 3.4. There are multiple simulations done for more statistical and reliable results and for every simulation, this random placement of incident area is made.

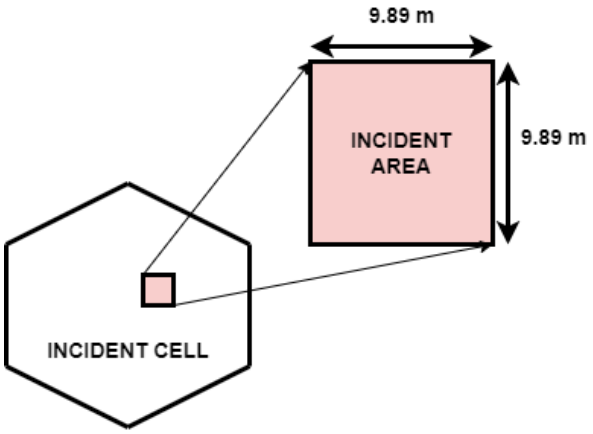

Figure 3.4: An incident cell visualization in the network [41].

Table 3.2: Traffic models.

Application		Service category	Session arrival process	Session duration (s)	File/ packet size (kB)	Packet arrival process
Broadband access everywhere		eMBB	Non-persistent with λ_{BB-DL}=100		2000 [1]	Bulk arrival
			Non-persistent with λ_{BB-UL}=40			
VR	Video content	URLLC	Non-persistent with λ_{VR}=40	5	1.25 [39]	Poisson arrival with λ_{VR} = 200 [39]
	Motion feedback data				0.5 [40]	Periodic with T_{VR} = 0.004 s [40]
Video surveillance (CCTV)		eMBB	Persistent with N_{VS} = 676 [6]		4.5 [1]	Periodic with T_{VS} = 0.036 s [1]
Sensors	Non-risk-sensitive measure-ments	mMTC	Persistent with N_{S-NRS} = 472920 [6]		0.2 [1]	Periodic with T_{S-NRS} = 60 s [1]
	Risk-sensitive measure-ments		Persistent with N_{S-RS} = 202680 [6]		0.2 [1]	Periodic with T_{S-RS} = 60 s [1]
	Sending emergency messages	URLLC	Persistent with N_{SE} = 25		0.032 [1]	Periodic with T_{SE} = 0.03 s [39]
Body camera		eMBB	Persistent with N_{BC} = 2		4.5 [1]	Periodic with T_{BC} = 0.036 s [1]
AR	Video content	URLLC	Persistent with N_{AR} = 2		1.25 [39]	Poisson arrival with λ_{AR} = 200 [39]
	Motion feedback data				0.5 [40]	Periodic with T_{AR} = 0.004 s [40]

3.4. Estimation of bandwidth

In this study, the channel bandwidth is calculated based on the average traffic volume of each application offered to the cell per second. In practice, together with average traffic volume, the performance requirements are also considered to evaluate the required channel bandwidth. The objective of this study is not to perfectly dimension the resources. Therefore, a rough estimation of channel bandwidth is made based on average traffic volume aspect. The average traffic volume in a cell per application excluding the applications arriving with an incident is shown in Table 3.3. In reality, the incident will not be always present in the network. Hence, for handling the unpredictable traffic

such as due to an incident and variability in inherent traffic, the service provider may likely slightly over-dimension the estimated channel bandwidth and 15% over-dimensioning is chosen for this study.

Table 3.3: Average traffic volume of each application in each cell in the network (without incident).

Application		**Average traffic volume/cell (Mbps) (DL)**	**Average traffic volume/cell (Mbps) (UL)**
Broadband access everywhere		28.07	11.22
VR	Video content	7.01	
	Motion feedback data		3.50
Video surveillance (CCTV)			11.85
Sensors	Non-risk-sensitive measurements		0.22
	Risk-sensitive measurements		0.09
Total average traffic volume (DL and UL)		35.08	26.88
Total average traffic volume		61.96	

The estimated channel bandwidth is calculated using the average spectral efficiency of 5G networks, which is defined as:

$$\text{Average spectral efficiency 5G} = \frac{\text{5G NR Throughput (bps)}}{\text{Channel bandwidth (Hz)}} \tag{3.6}$$

The average spectral efficiency for a dense-urban eMBB traffic is expected to be 7.8 bps/Hz for DL and 5.4 bps/Hz for UL [30]. This average spectral efficiency is defined based on the assumption of using techniques like single/multi-user MIMO beamforming. As these techniques are not modelled in this study, it is expected that the spectral efficiency which can be achieved in this simulation study will be lower.

The total estimated average traffic volume for DL and UL are 35.08 Mbps and 26.88 Mbps respectively, as shown in Table 3.3. Using (3.6) and the above-mentioned spectral efficiency, the minimum channel bandwidth required to serve the considered DL and UL traffic is estimated to be 4.49 MHz and 4.23 MHz respectively. Thus, the minimum total estimated channel bandwidth required is 8.72 MHz. After adding 15% to the estimated channel bandwidth, the final estimated channel bandwidth is 10.02 MHz.

As already mentioned, the spectral efficiency will be lower than in 5G deployments including advanced technologies and features which are excluded here and thus the channel bandwidth required will be higher than the estimated channel bandwidth. Therefore, simulations will be performed with channel bandwidths of 10, 15, 20 MHz or higher to determine the lowest bandwidth with which the applications experience satisfactory performance.

3.5. RAN Configurations

In this study, the RAN is configured with one or a set of RAN features, explained in previous chapter 2, to evaluate the impact of different features in achieving the target performance requirements of each application. The choices of the RAN configurations and their modelling aspects are explained in this section. Further, an emergency incident is modelled in the study, as explained in Section 1.2. To evaluate the impact of an incident in the network, the configurations are modeled for both

normal and incident scenario. For both normal and incident scenario, all the cells in the network are configured with same configuration under investigation. The RAN configurations are classified into three groups based on the three architectural choices, mentioned in Section 1.3 namely, non-sliced RAN, three RAN slices, each per service category and four RAN slices, each per service category and one slice for emergency service group.

3.5.1. Non-sliced RAN configurations

The RAN is configured with a number of numerologies, based on the concept of *flexible numerology*, to find the impact of each numerology on performance of the applications. The numerology value is limited based on the carrier frequency, as previously mentioned in Section 2.2. In this study, the carrier frequency f = 3.5 GHz is considered, which falls in the sub 7.125 GHz band or in the FR1. Hence, numerology 0,1 and 2 are considered in this study.

Assuming a non-sliced RAN architecture, RAN configurations with *BWPs* are considered, where the RAN is multiplexed with different numerologies in a non-sliced RAN architecture. The *number* of BWPs are decided based on the suitable numerology required to fulfill the diverse service requirements of the applications in the network. Based on the trade-off between latency and throughput, as explained in Section 2.2, the RAN is configured with two BWPs, one for the eMBB and mMTC traffic with throughput and no specific requirements, respectively, and another BWP for the URLLC traffic with latency requirement. The *numerologies* of the BWPs are decided based on the analysis obtained from the previous RAN configurations with different numerologies. The *resource split ratio* between the BWPs will be estimated, based on the resource utilization of the BWPs application, obtained from the simulations results of the previous RAN configurations with different numerologies.

Example: Consider the RAN configuration where the RAN is divided into two BWPs, configured with numerology 0 and 1, respectively, with the required middle guard band between BWPs, as shown in Figure 3.5. Now, consider the fractions of average resource utilization by the BWP_0 ap-

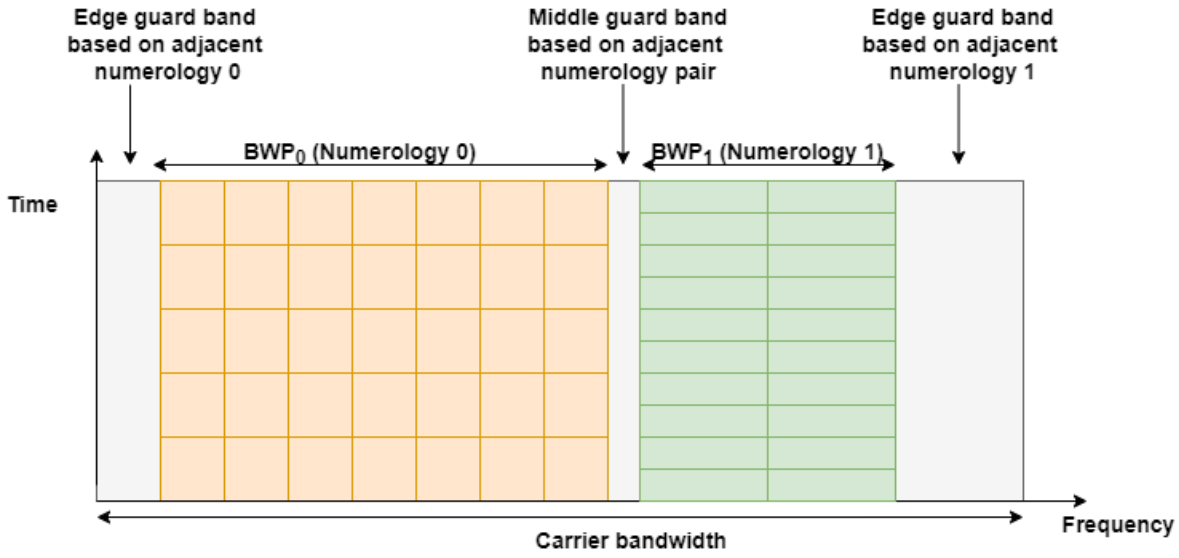

Figure 3.5: One visual example of RAN configured with two BWPs.

plications are d_0 and u_0 for handling the DL and UL traffic, respectively in the whole carrier and by BWP_1 applications are d_1 and u_1 for handling the DL and UL traffic, respectively. The fractions (d_0 and u_0) and (d_1 and u_1) are obtained from the simulation results of the RAN configurations with numerology 0 and 1, respectively. Then, the bandwidth utilized for handling the traffic of the BWP_0 and BWP_1 applications are calculated by multiplying the fraction of resource utilization by the SCS of the numerology of the BWPs because the BWPs are configured with different numerologies

which has different SCS. The d_0 and u_0 are multiplied with 15 kHz (SCS for numerology 0) and d_1 and u_1 are multiplied with 30 kHz (SCS for numerology 1). Then, the resource split ratios between the BWPs for both DL and UL slots are calculated because the DL and UL traffic load ratio between the BWPs are different. This BWP resource split ratio is calculated by taking the ratio of the DL and UL bandwidth utilization, respectively:

$$\text{BWPs resource split ratio (DL)} = \frac{\text{x} \times 15 \text{ kHz}}{\text{l} \times 30 \text{ kHz}} \tag{3.7}$$

$$\text{BWPs resource split ratio (UL)} = \frac{\text{y} \times 15 \text{ kHz}}{\text{m} \times 30 \text{ kHz}} \tag{3.8}$$

The resources utilized for the middle and edge guard bands are reduced from the total available resources in order to estimate the absolute size of the BWPs.

With division of RAN into BWPs, the resources can be utilized in a more efficient way by utilizing the idle resource sharing capability between BWPs, as explained in Section 2.6. The configurations considered with BWPs are modeled and simulated with and without idle resource sharing capability between BWPs. Another way of efficient resource utilization is explained in Section 2.5.2, by enabling mini-slot based scheduling. The RAN configurations are modeled by enabling each of the three mini-slot based scheduling schemes. Also, the RAN configurations with BWPs are simulated with mini-slot based scheduling to find the impact of combining these features on the performance of the applications.

For all the non-sliced RAN configurations, the DL and UL traffic ratio in the whole network is taken into account to find the TDD configuration. The TDD frame size for all the configurations in this study is chosen to be five. The DL and UL traffic ratio is determined by the average DL and UL traffic volume per cell in the network. This ratio is used to find the TDD configuration and special slot configuration, as explained in Section 2.4. Table 3.4 shows the calculated TDD configuration and the special slot configuration for both normal and incident scenarios, respectively.

Table 3.4: TDD configuration for non-sliced RAN configurations with normal and incident scenarios.

RAN configuration	**Average DL traffic volume/ cell (Mbps)**	**Average UL traffic volume/ cell (Mbps)**	**DL/UL traffic ratio of all applications**	**TDD configuration**	**Special slot (S) (DL:GP:UL)**
RAN configured with/without BWPs for normal scenario	35.08	26.88	1.30	D D S U U	10:2:2
RAN configured with/without BWPs for incident scenario	39.08	30.88	1.27	D D S U U	10:2:2

3.5.2. Sliced RAN configurations

Two slicing scenarios are considered in this study: (i) where the RAN is configured with RAN slices dedicated per service category and (ii) where the RAN is additionally configured with a RAN slice dedicated for emergency service group (customer) besides the other RAN slices per service category as in (i).

For these two slicing scenarios, for each considered RAN configuration, the RAN is configured with slice-specific choice of numerology. The TDD frame configuration for each slice is calculated based on the DL and UL traffic volume of slice-specific applications. Additionally, the number of

resources assigned to the slices will be determined based on the average resource utilization of the slice-specific applications, obtained from simulation results, for the slice-specific numerology, as explained above for the case of BWPs. The choice of RAN configurations and their modelling aspects for both slicing scenarios are explained in next subsections.

3.5.2.1. Slices per service category

In this scenario, the RAN is configured with three RAN slices, where each RAN slice is dedicated to a service category, as shown in Figure 3.6. The numerologies considered for configuration of each RAN slice are based on the trade-offs presented in Section 2.2 and they are shown in Table 3.7 and 3.8. For the URLLC slice, configurations are considered with both numerology 1 and 2, respectively and based on the comparison between the simulation results of these configurations with URLLC slice configured with numerology 1 and 2, the most suitable numerology will be selected.

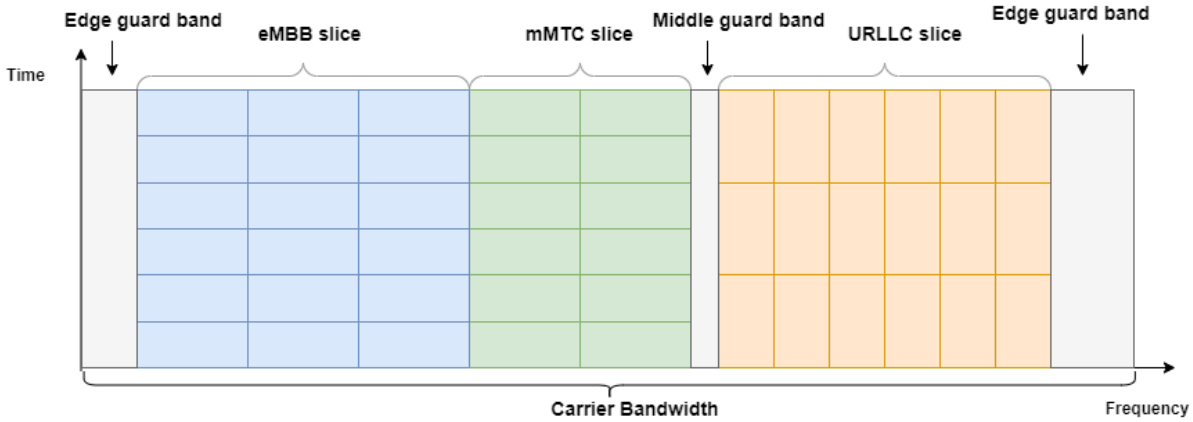

Figure 3.6: One visual example of configuration with three slices, each per service category.

Additionally, and as discussed in Section 2.4, each RAN slice can be individually configured with a different TDD configuration based on the DL/UL traffic ratio of each slice. The slice-specific applications with the corresponding aggregated traffic load and DL/UL traffic ratio per slice for the normal and incident scenarios, are shown in Table 3.5 and 3.6, respectively.

Table 3.5: Average DL and UL aggregated traffic load per slice for configurations with three slices, for normal scenario.

Slice	**Applications**	**Average DL traffic volume/cell (Mbps)**	**Average UL traffic volume/cell (Mbps)**	**DL/UL traffic ratio**
eMBB	Broadband access everywhere, video surveillance (CCTV)	28.07	23.07	1.21
mMTC	Sensors monitoring non-risk-sensitive measurements, sensors monitoring risk-sensitive measurements	0	0.31	0
URLLC	VR (Motion-feedback data), VR (Video Content)	7.01	3.50	2.00

Based on the aggregated average traffic volume per slice and the corresponding DL/UL ratio as shown in Table 3.5, the TDD configuration for each RAN slice is derived, as explained in Section 2.4 which is shown in Table 3.7 for normal scenario. The TDD frame size for all the slices is chosen to be five.

Table 3.6: Average DL and UL aggregated traffic load per slice for configurations with three slices, for incident scenario.

Slice	Applications	Average DL traffic volume/cell (Mbps)	Average UL traffic volume/cell (Mbps)	DL/UL traffic ratio
eMBB	Broadband access everywhere, video surveillance (CCTV), body camera	28.07	25.07	1.11
mMTC	Sensors monitoring for non-risk-sensitive measurements, sensors monitoring for risk-sensitive measurements	0	0.31	0
URLLC	VR (Motion-feedback data), VR (Video Content), AR (Motion-feedback data), AR (Video Content), sensors sending emergency messages	11.01	5.50	2.00

Table 3.7: Numerologies and TDD configuration per slice for configurations with three slices, for normal scenario.

Slice	Considered numerologies	DL/UL ratio	TDD configuration	Special slot (S) (DL:GP:UL)
eMBB	0	1.21	D D S U U	9:2:3
mMTC	0	Only UL traffic but DL symbols for control messages	S U U U U	10:2:2
URLLC	1,2	2.00	D D D S U	3:2:9

For the scenario with an incident, a new TDD configuration is again derived for each slice as the traffic load per slice changes, when the incident occurs. The new TDD configuration is considered for all the cells in the network, as the incident occurs in a randomly selected cell during the simulation. Table 3.8 show the adapted TDD configuration for each slice for the incident scenario.

Table 3.8: Numerologies and TDD configuration per slice for configurations with three slices, for incident scenario.

Slice	Considered numerologies	DL/UL ratio	TDD configuration	Special slot (S) (DL:GP:UL)
eMBB	0	1.11	D D S U U	8:2:4
mMTC	0	Only UL traffic but DL symbols for control messages	S U U U U	10:2:2
URLLC	1,2	2.00	D D D S U	3:2:9

Like the configurations with BWPs, the option of idle resource sharing between slices is considered. Similarly, the mini-slot based scheduling feature is combined with three slice configurations. Among the three mini-slot based scheduling schemes, described in Section 2.5.2, the most appropriate scheme will be selected from the simulation results which will be obtained from configurations where different mini-slot based scheduling schemes are considered.

3.5.2.2. Slice per service category and emergency service group (customer)

Another approach of configuring the RAN is to have a separate RAN slice for a vertical customer [16]. In this study, the emergency service group is identified as a vertical customer with mixed traffic

requirements as mentioned in Section 1.2. The RAN is configured with a separate RAN slice for the customer in the whole network, together with the three RAN slices per service category. The slice-specific applications with the corresponding aggregated traffic load and DL/UL traffic ratio per slice for the normal and incident scenarios, are shown in Table 3.9 and 3.10, respectively.

Table 3.9: Average DL and UL aggregated traffic load per slice for configurations with four slices, for normal scenario.

Slice	**Applications**	**Average DL traffic volume/cell (Mbps)**	**Average UL traffic volume/cell (Mbps)**	**DL/UL traffic ratio**
eMBB	Broadband access everywhere	28.07	11.22	2.50
mMTC	Sensors monitoring non-risk-sensitive measurements	0	0.22	0
URLLC	VR (Motion-feedback data), VR (Video content)	7.01	3.50	2.00
Customer	Video surveillance (CCTV), sensors monitoring risk-sensitive measurements	0	11.94	0

Table 3.10: Average DL and UL aggregated traffic load per slice for configurations with four slices, for incident scenario.

Slice	**Applications**	**Average DL traffic volume/cell (Mbps)**	**Average UL traffic volume/cell (Mbps)**	**DL/UL traffic ratio**
eMBB	Broadband access everywhere	28.07	11.00	2.50
mMTC	Sensors monitoring non-risk-sensitive measurements	0	0.22	0
URLLC	VR (Motion-feedback data), VR (Video content)	7.01	3.50	2.00
Customer	Video surveillance (CCTV), sensors monitoring risk-sensitive measurements, body camera, AR (Video content), AR (Motion feedback data), sensors sending emergency messages	4	15.94	0.25

The numerologies considered for configuration of each RAN slice are based on the trade-offs presented in Section 2.2, are shown in Table 3.11 and 3.12. For URLLC slice, the numerology selection will be done based on the simulation results of configurations with three slices. For customer slice, numerology 0 is chosen because of associated eMBB and mMTC applications, for the normal scenario. Figure 3.7 shows an example of one such RAN configuration for normal scenario with URLLC slice configured with numerology 1.

For incident scenario, configurations with all three numerologies are considered for the customer slice because of associated mixed application requirements. Based on the simulation results, the appropriate numerology will be selected for the customer slice. If from the simulation results, no one suitable numerology is found for customer slice, then another configuration will be considered with division of customer slice into required number of BWPs.

Additionally, and similar to other slicing scenario, each RAN slice can be configured with a different TDD configuration, based on the slice-specific DL/UL traffic ratio. The TDD configurations

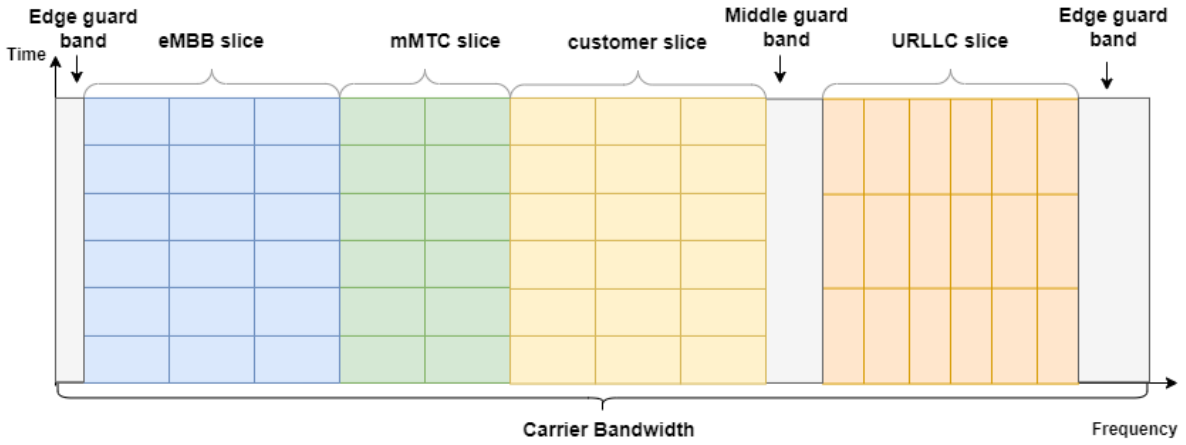

Figure 3.7: One visual example of configuration with four slices, each per service category and one for the customer, for normal scenario.

for each RAN slice are shown in Table 3.11, for normal scenario. For incident scenario, a different TDD configuration is considered, as also explained in slicing scenario with three slices. Table 3.12 show the adapted TDD configurations for each slice for the incident scenario.

Table 3.11: Numerologies and TDD configuration per slice for configurations with four slices, for normal scenario.

Slice	Considered numerologies	DL/UL ratio	TDD configuration	Special slot (S) (DL:GP:UL)
eMBB	0	2.5	D D D S U	7:2:5
mMTC	0	Only UL traffic but DL symbols for control messages	S U U U U	10:2:2
URLLC	1,2	2.00	D D D S U	3:2:9
customer	0	Only UL traffic but DL symbols for control messages	S U U U U	10:2:2

For the configurations with four slices, the decision of whether to enable the idle resource sharing between slices will be taken from the analysis obtained for configurations with BWPs and configurations with three slices. Similarly, the choice of whether to schedule using mini-slots and the most beneficial mini-slot based scheduling scheme will be selected from the analysis obtained from configurations with BWPs and three slices.

3.6. Data Transmission

The DL and UL transmission procedures, the processing delays and the physical channels used for a transmission are briefly described in Section 3.6.1 and 3.6.2, respectively.

3.6.1. DL Transmission

The UE periodically sends a Channel State Information (CSI) report on the Physical Uplink Control Channel (PUCCH) to the gNB which comprises a CQI. The UE reports a CQI per sub-band, where a sub-band is a subset of the total bandwidth. The size of the sub-band depends on the total number of PRBs in the total bandwidth [42]. Thus, the size of the sub-band varies with different numerologies. The UE then estimates the *Signal-to-Interference-plus-Noise Ratio (SINR)* of each PRB in the sub-band and from that derives the CQI for the given sub-band. The SINR is estimated for every

Table 3.12: Numerologies and TDD configuration per slice for configurations with four slices, for incident scenario.

Slice	Considered numerologies	DL/UL ratio	TDD Configuration	Special slot (S) (DL:GP:UL)
eMBB	0	2.5	D D D S U	7:2:5
mMTC	0	Only UL traffic but DL symbols for control messages	S U U U U	10:2:2
URLLC	1,2	2.00	D D D S U	3:2:9
customer	0,1,2	0.25	D S U U U	0:2:12

scheduling slot as:

$$SINR^{PRB} = \frac{S^{PRB}}{N_0 \times NF_{UE} \times B^{PRB} + I^{PRB}} \tag{3.9}$$

where, S^{PRB} is the received signal power from the serving BS, N_0 is the thermal noise density, NF_{UE} refers to the receiver noise figure, B^{PRB} is the PRB bandwidth and I^{PRB} is the aggregation of all the received power from the non-serving cells in the network contributing to the interference. Unlike regular slot based scheduling, where the SINR is estimated per scheduling slot, with mini-slot based scheduling, the SINR is estimated symbol-wise. If a mini-slot of 7 symbols is used for transmission, then the SINR is estimated for 7 symbols of the regular slot. Similarly, for S slot, the SINR is estimated separately for DL and UL symbols based on the S slot configuration.

Subsequently, for each $SINR^{PRB}$, a Mutual Information (MI) metric is calculated using the Mutual Information-Effective SINR Mapping (MI-ESM) algorithm [43]. The MI values, derived for all PRBs in the sub-band, are then averaged. The averaged MI value is then mapped to find the effective SINR for the sub-band again using the MI-ESM curves. The curves are shown in Appendix. The effective SINR and the target Block Error Rate (BLER) are then used to find the appropriate CQI based on the BLER curves for an Additive White Gaussian Noise (AWGN) channel. The target BLER is set to not exceed 0.001% for URLLC transmisisons and 0.1% for eMBB and mMTC transmissions. The BLER curves indicate, for each CQI value, the maximum SINR value to achieve the target BLER as also shown in Appendix. The BLER curves are derived using the Vienna 5G Link Level Simulator [44].

The UE reports the CQI based on the measured SINR and hence indicates the *MCS* to be used for transmission to the BS, as detailed in [42]. As the channel quality and transmission activity of BSs changes over time and frequency, the reported CQI can be inaccurate due to a delay between the moment the BS uses the reported CQI and the time when the SINR is measured by the UE. To overcome the effects of the outdatedness in terms of excessive BLERs, an Outer-Loop Link Adaptation technique (OLLA) is used. The OLLA technique effectively increases or decreases the MCS (as indicated by the reported CQI) by an adaptively tuned offset, to prevent the realized BLER from exceeding beyond its required level, as also detailed in [45].

The gNB then needs to schedule the PRBs to the UEs to transmit their data. For *scheduling* the PRBs, the bit rate per PRB corresponding to the MCS is used by the *packet scheduler* to find which PRBs to be assigned to which UE. This is done by using the scheduling metric M of the packet scheduler. Considering the diverse QoS requirements of the chosen scenario in this study, the M-LWDF packet scheduler is implemented. The M-LWDF scheduler can serve eMBB, mMTC and URLLC traffic as it takes into account both latency-oriented and channel-adaptive aspects. For eMBB and mMTC traffic, the M-LWDF scheduler follows the principles of the PF scheduler. The PF scheduler takes into account the currently attainable bit rate and the average experienced user bit rate and distribute the resources such that on average, all the UEs get equal share of resources over time. The

scheduler metric M used for eMBB and mMTC flows is given in [33] as:

$$M_{M-LWDF,i}(t,f) = \frac{R_i(t,f)}{\bar{R}_i(t-1)} \quad \forall i \in \text{eMBB and mMTC flows} \tag{3.10}$$

where, $R_i(t,f)$ is the attainable bit rate at time t and PRB f and $\bar{R}_i(t)$ is the experienced bit rate of the active flow i till the current TTI t [42]. The M-LWDF scheduler serves the URLLC traffic with a weighted version of the PF scheduler metric [42] as shown in (3.10):

$$M_{M-LWDF,i}(t,f) = -\frac{W_i(t)\log\delta_i}{\tau_i} \times \frac{R_i(t,f)}{\bar{R}_i(t-1)} \quad \forall i \in \text{URLLC flows} \tag{3.11}$$

where $W_i(t)$ is the delay from time of arrival of the active flow, δ_i denotes the maximum allowed packet drop rate for an active flow i and τ_i is the latency budget for the active flow i in the RAN [46]. The scheduler assigns those PRBs to the UE which obtain the highest scheduler metric values.

Based on the aggregate number of PRBs assigned to the UE by the scheduler, the effective SINR of the assigned PRBs is calculated by again using the MI-ESM algorithm. The effective SINR and the target BLER is again used to find the CQI of the sub-band of assigned PRBs. The CQI indicates the MCS and code rate to be used for transmission, as explained in [42]. The MCS, code rate and the number of assigned PRBs are used to calculate the Transport Block Size (TBS) [42]. The TBS is the data scheduled in one transmission to a given UE. The gNB then transmits the data to the UE on the Physical Downlink Shared Channel (PDSCH). Depending on the TDD configuration, there can be delay between the DL allocation and transmission because of the waiting time for the next DL slot.

Upon reception of the data, the UE processes the received data and sends an Acknowledgement (ACK)/ Negative Achowledgement (NACK) on the PUCCH in order to indicate the correct or wrongful reception of the data. The processing time between the DL data reception and the feedback depends on the configured numerology and the UE capability [42]. There can be an additional waiting time for an UL slot to send the feedback, depending on the TDD configuration. If the UE sends a negative acknowledgment, the gNB prepares a retransmission if it is possible to be received by the UE within the required latency budget. The total time between the NACK transmission and the DL data retransmission is the latency required to decode the feedback and to prepare the data for retransmission. The UL and DL transmission time, as well as the processing time for transmission and reception of data at the gNB, is equal to one slot [33].

Figure 3.8 shows the procedure, overall latencies and the physical channels used for a DL transmission with one retransmission [24]. The delay occurred because of waiting time for UL and DL channels is not marked in the Figure 3.8. Moreover, the physical control channels PUCCH and Physical Downlink Control Channel (PDCCH) are not explicitly modelled in this study, but the control messages sent together with the data on the shared channel are taken into account in terms of the resources they consume and the delay they impose. The resources required for the control messages are reduced while calculating the TBS.

3.6.2. UL Transmission

The UE requests an UL transmission by sending a Scheduling Request (SR) which is a control message sent by the UE on the PUCCH to the gNB. The gNB processes the request and sends an UL grant to the UE which includes the parameter K2, which is the processing delay between the UL grant and the UL transmission [42], the length of the scheduled transmission, the start symbol of the UL transmission, the MCS and the transmit power for transmission. The gNB derives the transmit power to be used by the UE for the UL data transmission on the PUSCH using:

$$P_{PUSCH} = min(P_{max}, P_o + 10log_{10}(2^{\mu} \times (M)) + \alpha \times PL) \tag{3.12}$$

where P_{max} is the maximum transmit power of the UE, P_o is the UE/cell specific power offset signalled by radio resource control, μ is the numerology value with which the RAN is configured, M is

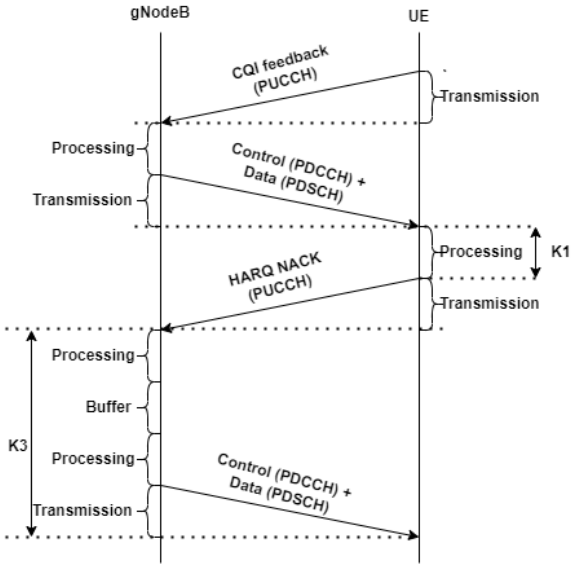

Figure 3.8: Procedure and physical channels used for DL transmission with one retransmission [24].

the number of assigned PRBs to the UE, α is the fractional path loss compensation factor and PL refers to the path loss. The parameter P_o and α are taken as -60 dBm and 0.5, respectively, based on [47]. These values are expected to maximize the spectral efficiency in the network. The packet scheduler assigns UE the minimum required PRBs f based on the scheduler metric M, as explained in DL procedure. To derive the scheduler metric for each PRB for the scheduling process, first the SINR per PRB is calculated, which is given as:

$$SINR^{PRB} = \frac{S^{PRB}}{N_0 \times NF_{BS} \times B^{PRB} + I^{PRB}} \tag{3.13}$$

where, S^{PRB} is the received power from the UE, N_0 is the thermal noise density, NF_{BS} is the receiver noise figure in linear units, B^{PRB} is the PRB bandwidth and I^{PRB} is the aggregation of all the received power from the other UEs served by the same or other gNBs in the network contributing to the interference. The received power of the UE is derived by using the open loop power spectral density defined as the Sounding Reference Signal (SRS) transmit power, given as:

$$P_{SRS}^{PRB} = min(P_{max}, P_o + 10log_{10}(2^{\mu} \times (M)) + \alpha \times PL) \tag{3.14}$$

The UE periodically sends an SRS signal which is an UL reference signal to the gNB. The UE sends the SRS per sub-band. The size of the sub-band are same as explained in DL procedure. After finding the SINR per PRB, the gNB derives the effective SINR per sub-band using the MI-ESM algorithm and derives the MCS and assigned number of PRBs for the requested UL transmission by the UE, similarly as explained in DL procedure. The gNB then sends the MCS, transmit power for transmission, the parameter K2 and the assigned PRBs to the UE as part of UL grant.

The UE processes the received UL grant and prepares the data based on the information received in the UL grant. The processing time between the UL grant reception and the UL data transmission is given by parameter K2. The processing time K2 depends on the configured numerology and the UE capability [42]. Upon reception of the data, the gNB processes the received UL data. In case of wrongful reception of data, the gNB sends a NACK and a retransmission request to the UE. The UE processes the request and prepares the UL data for retransmission only if the data can be received within the required latency budget. The UL and DL transmission time, as well as the

processing time for transmission and reception of data at the gNB, is equal to one slot [33]. The TBS for UL transmission is calculated, as explained in the DL procedure.

Figure 3.9 shows the UL transmission procedure and the relevant latencies and the physical channels with one retransmission. The latency occurred because of waiting time for UL and DL channels is not marked in the Figure 3.9. The physical control channels PUCCH and PDCCH are not explicitly modelled in this study, but the control messages sent together with the data on the shared channel are taken into account in terms of resources. The resources required for the control messages are reduced while calculating the TBS.

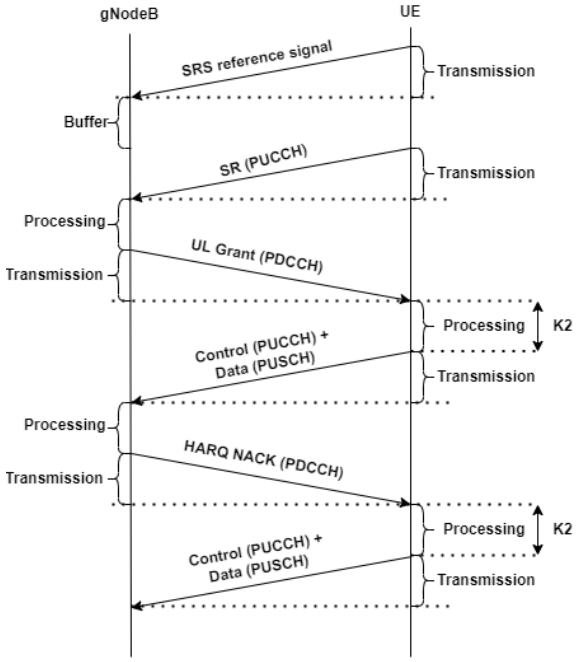

Figure 3.9: Procedure and physical channels used for UL transmission with one retransmission.

3.7. Simulation Flow

System-level simulations are performed for evaluating the features explained in Chapter 2, for all the choices of RAN configurations explained in Section 3.5 for both the normal and incident scenarios. A high-level overview of the simulation flow is presented in Figure 3.10.

The simulations are initialized by generating all the users corresponding to each application, according to the considered traffic models explained in Section 3.3. All the users per applications are then distributed in the network. Then, the timeline and the channels are created for all persistent and non-persistent users. As the network consists of 57 cells, the generated traffic is assigned per cell. Based on the RAN configuration under investigation, the traffic is then assigned to the BWP or slice.

The corresponding TDD configuration, for the RAN configuration under investigation, is checked and according to the type of scheduling slot (DL, UL or S), the relevant traffic is scheduled in the given slot. The packet scheduler assigns the resources to the active users, in the given scheduling slot, based on the scheduler metric and the type of scheduling scheme enabled in the given configuration, e.g. including the use of mini-slots. If the configuration consists of slices or BWPs, then the resources are shared between the slices or BWPs, if the sharing capability is enabled in the given configuration. The transmission parameters for each block transmission are then derived, the SINR is derived, and a correspondingly biased coin is flipped to determine whether the

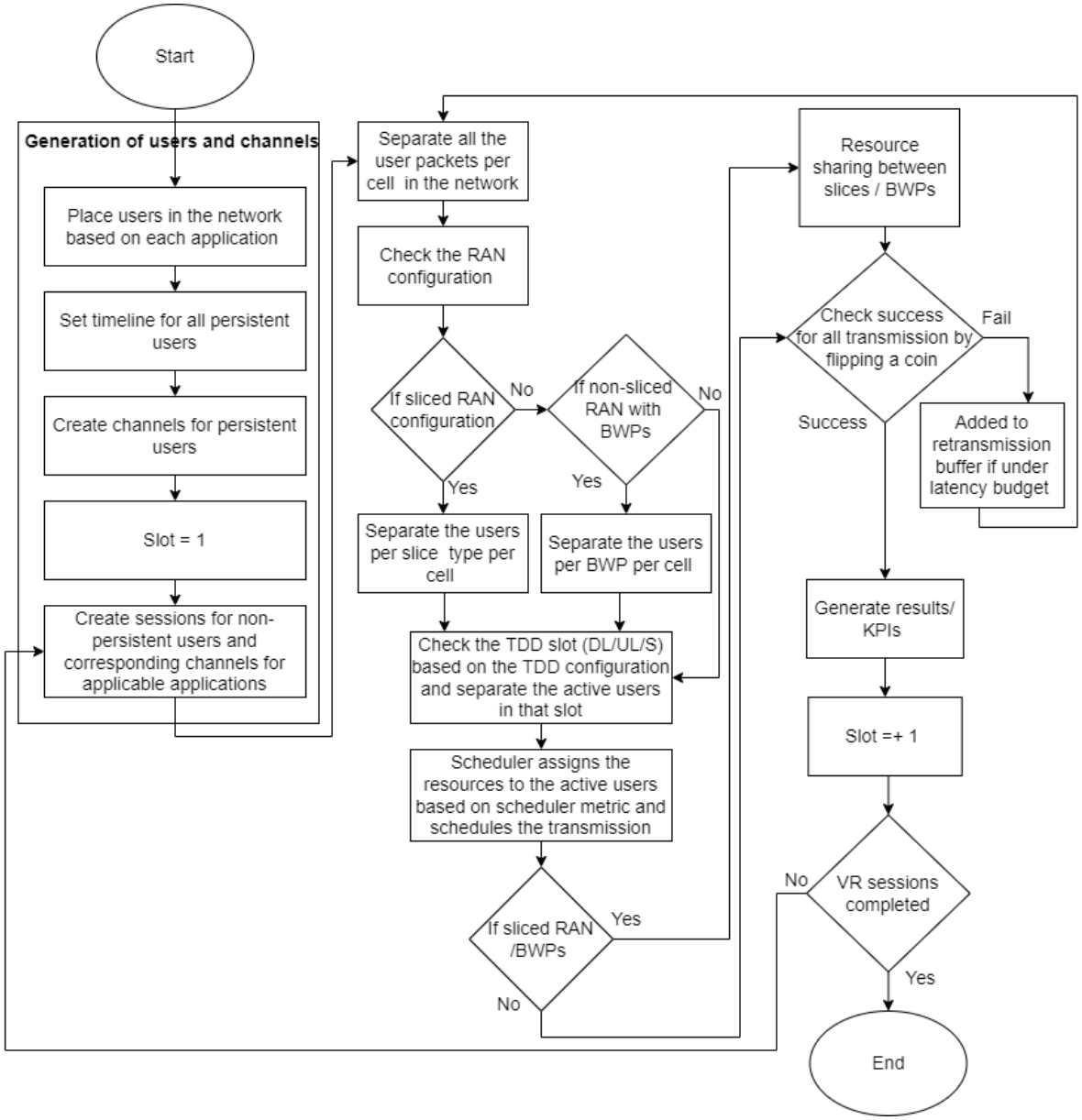

Figure 3.10: Outline of the simulation flow.

block is received correctly. In case of successful transmission, the results are generated for all the packets or part of packet transmitted in that slot. In case, the transmission is unsuccessful, the corresponding packets are appended back to the retransmission buffer only if, another transmission can be sent within the latency budget (if applicable such as for URLLC transmissions). Otherwise, the packets are considered lost. The next scheduling slot is then checked, and the same simulation cycle continues until the end of the simulation period. The simulation period is derived considering several factors explained in the next paragraph.

The KPI values for each application, as explained further in the next Chapter, are derived by analyzing the transmissions done after a warm-up period, which is the duration after which the traffic load in the network is stable. Additionally, the results from only the inner cells of the network, as shown in Figure 3.1, are considered because the outer cells will experience less interference in comparison with inner cells, which is unrealistic, since in reality the network area is not confined only to the considered smart city scenario. Among all the applications considered in this study,

the minimum number of users that arrive in the network per second is for VR applications, for the normal scenario. Therefore, the duration of simulation is dependent on completion of one session of 5 seconds for all the VR devices that arrive in a second after the warm-up period in the network.

For each RAN configuration, multiple simulations are performed using distinct random seeds to achieve more statistically reliable results. For every simulation run, the locations of the users are generated randomly. Also, for each cell-user pair, the multipath fading trace and the starting slot within the trace is selected randomly. The trace is wraparound once the end of the trace is reached. This way, for every simulation run, each user will have a different location and different multipath effect towards each cell, which will result in a more statistically reliable evaluation of smart city scenario with these type of applications.

4 Assessment of 5G Key enablers

In this chapter, the assessment of the 5G-RAN technological features explained in Chapter 2 in achieving the diverse service requirements of the considered applications are presented. The KPIs which are used to assess the RAN features are explained in Section 4.1. The analysis done to estimate the required carrier bandwidth is presented in Section 4.2. Section 4.3 presents the performance of the considered RAN features under different RAN configurations for the normal scenario, followed by the assessment for the incident scenario in Section 4.4. The best-performing RAN configurations, among the considered configurations for normal and incident scenarios, are provided in Section 4.5.

4.1. Key Performance Indicators (KPIs)

There are three KPIs considered in this study based on the service requirements of the considered applications. The definitions and process of obtaining each of these KPIs are explained below:

- **5th throughput percentile:** For the eMBB applications, i.e. broadband access everywhere, video surveillance and body camera, the focus is on the throughput metric. The throughput of each device is defined as the average throughput of all the packet transmissions between the gNodeB and the device during the simulation period. Based on the requirements mentioned in Section 1.2 for each application, 95% of the devices should experience higher than the respective target throughput and therefore the 5th throughput percentile is considered as the KPI of relevance. The target throughput is different for each eMBB application, and it is presented in Section 1.2. This KPI is derived for each application by first calculating the Cumulative Distribution Function (CDF) of the throughput of each device, of the corresponding application, and then the 5th percentile is selected.

- **Percentage of devices experiencing a certain reliability:** For the URLLC applications, i.e. VR, AR and sensors sending emergency messages, the focus is on the latency of a transmission with a certain reliability. The latency is the time interval between the transmission and reception of packets at the destination, and the reliability is defined as the fraction of successful transmissions within the target latency budget. Similar to the eMBB applications, 95% of the devices per application should experience the target reliability. The target reliability requirement is different for each application, and it is presented in Section 1.2. To derive this KPI, the percentage of the successful transmissions within the latency budget is calculated for each device of the concerned application. Then, the percentage of devices that achieve the required target reliability as per the application is calculated.

- **95th latency percentile:** For the mMTC applications, i.e. sensors monitoring risk-sensitive and non-risk sensitive measurements, even though there are no strict throughput or latency requirements identified, this study focuses on the latency of the packet transmission to assess the impact of the RAN features on this type of applications. The latency metric is chosen, instead of the throughput metric, because the devices related to the URLLC application "sensors sending emergency messages" are part of the mMTC application "sensors monitoring risk-sensitive measurement" until an emergency incident occurs. Also, in-line with the other applications, the latency experienced by 95% of the devices per application is assessed. Hence, the KPI for the mMTC applications is the 95th latency percentile, which is obtained by first calculating the CDF of the latency of the packet transmissions by all the devices and then the 95th percentile is selected.

Table 4.1 shows the summary of the relevant KPIs for each application and their corresponding target values. Further, to find statistical accuracy of each calculated KPI value, the relevant Confidence Interval (CI) is calculated. The CI is calculated using a number of KPI values, each obtained over a number of independent simulations, as explained in [48].

Table 4.1: Relevant KPIs and target values for each application.

Application	Scenario type	Service category	Relevant KPI	KPI target value
Broadband access everywhere	Normal	eMBB	5^{th} throughput percentile	50 Mbps (DL) and 12.5 Mbps (UL)
VR (Video content and motion-feedback data)	Normal	URLLC	Percentage of devices experiencing a certain reliability	95% of devices with 96% reliability for both DL and UL
Video surveillance (CCTV)	Normal	eMBB	5^{th} throughput percentile	25 Mbps
Sensors monitoring risk-sensitive and non-risk-sensitive measurements	Normal	mMTC	95^{th} latency percentile	-
Body camera	Incident	eMBB	5^{th} throughput percentile	25 Mbps
Sensors sending emergency messages	Incident	URLLC	Percentage of devices experiencing a certain reliability	95% of devices with 99.9% reliability
AR (Video content and motion-feedback data)	Incident	URLLC	Percentage of devices experiencing a certain reliability	95% of devices with 96% reliability for both DL and UL

4.2. Total bandwidth estimation

In Section 3.4, the carrier bandwidth required for this study is estimated to be 10 MHz based on the considered traffic load and the average 5G spectral efficiency, but it is likely that 10 MHz is not enough. This is because, the average spectral efficiency assumes the use of advanced techniques like MIMO beamforming, which is not modelled in this study. Also, to handle the inherent traffic variability and the unpredictable traffic i.e. incident-based traffic, some extra resources are reserved.

First, multiple simulations are performed with the previously estimated carrier bandwidth of 10 MHz to verify that a higher carrier bandwidth will be necessary. The DL and UL resource utilization per cell is averaged over these simulations, and the Probability Density Function (PDF) of the averaged resource utilization per cell is calculated and shown in Figure 4.1. Figure 4.1 shows that the average resource utilization per cell is likely to be in between 75% to 100% for both the DL and UL channels. Such a high resource utilization can significantly slow down the response time as well as introduce high interference, which results in performance degradation of the applications. Also, it is calculated that around 40% of the cells are likely to have 90% or higher resource utilization, which also reflects the need for more resources and thus a wider carrier bandwidth.

To find the required carrier bandwidth and a spectrum deployment granularity of 5 MHz [49], simulations are performed with a 15 MHz carrier bandwidth for the scenario where the RAN is only configured with numerology 0. This is, most likely, not an optimal configuration, but since this step of estimating the required carrier bandwidth is done only once, the most basic configuration

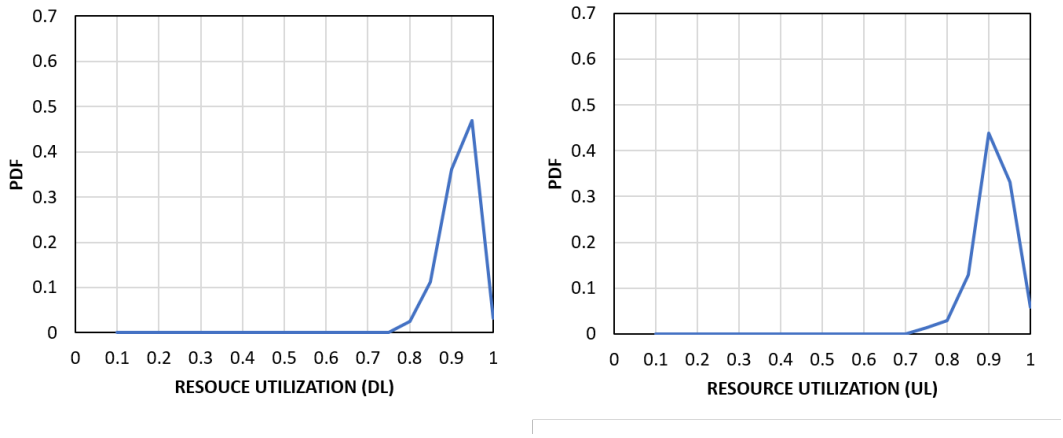

Figure 4.1: PDF of the DL and UL average resource utilization per cell with a 10 MHz carrier bandwidth for normal scenario.

is chosen, similar to 4G. The PDF of the DL and UL average resource utilization per cell is shown in Figure 4.2. It is observed in Figure 4.2 that the average resource utilization per cell is likely to be in between 45% and 75%, with a high probability of 65% for both the DL and UL channels. Considering the extra required resources, as mentioned in Section 3.4 and also in line with the typical operational deployment, the 65% of resource utilization is considered acceptable and the carrier bandwidth of 15 MHz is thus considered for the analysis to follow.

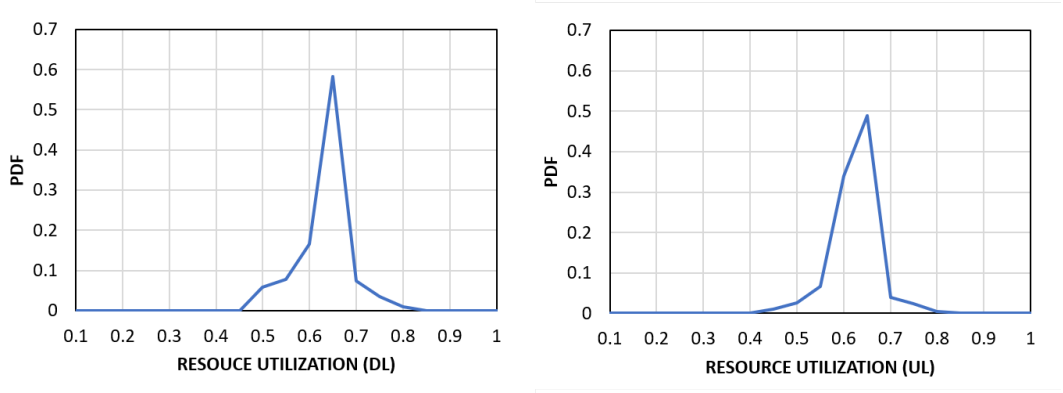

Figure 4.2: PDF of the average DL and UL average resource utilization per cell with a 15 MHz carrier bandwidth.

4.3. Assessment of 5G-RAN features for normal and incident scenario

This section presents the impact of the considered RAN features in achieving the considered application requirements. This is done by simulating and analyzing the results of different RAN configurations, where each configuration consists of a set of RAN features. In the graphical representations of the results, the notations are used where each notation indicates the enabled features in the configuration, as presented in Table 4.2. The notation is defined by combining the notation of the enabled features by an underscore. For example, when the RAN is configured with numerology 0 (notation: N0) and basic mini-slot based scheduling is enabled (notation: M), the notation of the configuration is given by N0_M. Also, in the graphical representation of the results, the applied notations are used to distinguish the results from different applications, the notations are shown in Table 4.3.

Considering the three architectural choices explained in Section 1.3, a number of RAN configu-

Table 4.2: The notation used in the graphical representations of each RAN feature.

RAN feature	Notation
Flexible numerology	NX (N = numerology, X = numerology value)
Basic mini-slot based scheduling	M
Non-pre-emptive mini-slot based scheduling	NP
Pre-emptive mini-slot based scheduling	P
Bandwidth Parts	XB (Y) (X= number of BWPs, B = use of BWP Y = numerologies used per BWP)
Slicing	XS (Y) (X = number of slices, S = use of slicing, Y = numerologies used per slice)
Resource sharing	RS
No middle guard band	NG

Table 4.3: The notation used in the graphical representation for each application.

Application	Notation
Broadband access everywhere	BB
VR (Video content and motion feedback data)	VR
Video surveillance (CCTV)	VS
Sensors for risk and non-risk sensitive measurements	S
Sensors sending emergency messages	SE
Body camera	BC
AR (Video content and motion feedback data)	AR

rations are considered per architecture presented as: a) non-sliced RAN configurations, b) RAN configurations with three slices, one per service category and c) RAN configurations with four slices, one per service category and one for the emergency service group (customer). First, the assessment of the considered RAN configurations per architectural choice on the performance of the applications is performed for the normal scenario (without an incident). The best-performing configurations for the normal scenario under each architectural choice is then assessed for a scenario with an incident.

4.3.1. Non-sliced RAN configurations

First, the RAN is configured with numerologies 0, 1 and 2 in different RAN configurations, respectively over the whole carrier, to find the impact of each numerology on the performance of each application. Then, different types of mini-slot based scheduling are considered to analyze and assess the benefits and/or losses of each type. Based on the obtained results from the numerology investigation, the number of BWPs and the suitable numerology of each BWP are configured and simulations are performed to study the performance of the BWPs, for scenarios with and without resource sharing between the BWPs. In case, the results obtained from the configuration where suitable numerology for each BWP is used, does not yield good performance, a new configuration is considered with use of sub-optimal numerology for BWPs. Finally, the configuration that combines BWPs with appropriate numerology per BWP, resource sharing and mini-slot based scheduling is considered to investigate the performance of the applications with combining all possible features

in a non-sliced RAN. The TDD configuration used in all considered non-sliced RAN configurations is the same and it is determined based on the DL to UL traffic ratio in the network, which is presented in Section 3.5.1. Also, the M-LWDF packet scheduler is considered for all the considered configurations.

4.3.1.1. Flexible numerology

The assessment of flexible numerology is made by configuring the whole carrier bandwidth with low to high numerologies i.e. 0, 1 and 2, respectively in three different simulation experiment and then by analyzing the obtained KPI values for each application for each configuration. The results are illustrated in Figure 4.3 with the target KPI values for each application, except for the mMTC application, indicated by the dashed lines. Specifically, Figure 4.3.a shows that for the URLLC application (VR), the highest numerology results in better performance than the lower numerologies, in the sense of maximizing the percentage of VR sessions satisfying the reliability requirement. This is because of two reasons: a) Packets are scheduled faster as the slot duration reduces with an increase in numerology, as mentioned in Section 2.2. b) The transmission time intervals are reduced with an increase in numerology, thus the actual transmission time is reduced. With both the improvements, the likelihood that the latency requirement is satisfied is increased with higher numerology.

On the other hand, it is observed in Figure 4.3.b and Figure 4.3.c that for the eMBB (BB and VS) and mMTC (S) applications, with the increase in numerology, the 5^{th} throughput percentile decreases and the 95^{th} delay percentile increases, respectively. This is because the increase of numerology reduces the total number of PRBs in a given bandwidth, which then leads to lower gains from frequency-selective channel-adaptive scheduling, in comparison with lower numerologies. Thus, the trade-off between the latency and throughput as qualitatively mentioned in Section 2.2 is proven to be true with the acquired results.

The target KPI values are met for two of the eMBB applications (BB (DL) and VS) and almost for BB(UL) application, only for the configuration with numerology 0. The target values are not achieved for any other applications, regardless of the numerology. Based on the analysis of these three configurations with low to high numerology, it is concluded that for URLLC applications, configuring the carrier with numerology 2 is beneficial and for the eMBB and the mMTC applications, numerology 0 is beneficial.

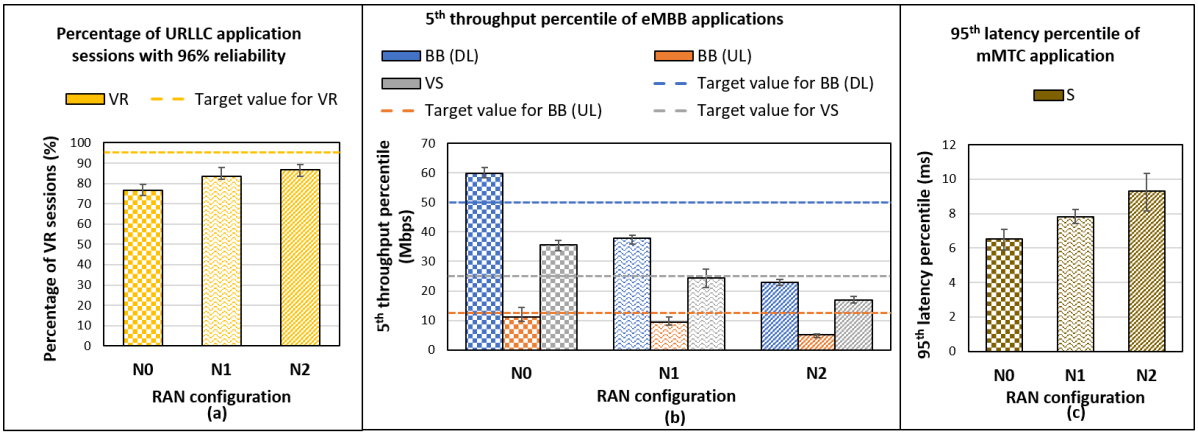

Figure 4.3: KPI values obtained by configuring the RAN with numerologies 0,1 and 2.

4.3.1.2. Mini-slot based scheduling

This subsection presents the impact of the three mini-slot based scheduling schemes explained in Section 2.5.2 on the performance of the applications. For a fair comparison, the RAN (whole carrier)

is configured with numerology 0 for all three schemes. The results obtained for these three RAN configurations are shown in Figure 4.4, where the three configurations are represented as N0_M, N0_NP and N0_P. Figure 4.4 also shows the previously obtained results with numerology 0, labeled as N_0, to allow a comparison between the regular slot-based scheduling and the mini-slot-based scheduling schemes.

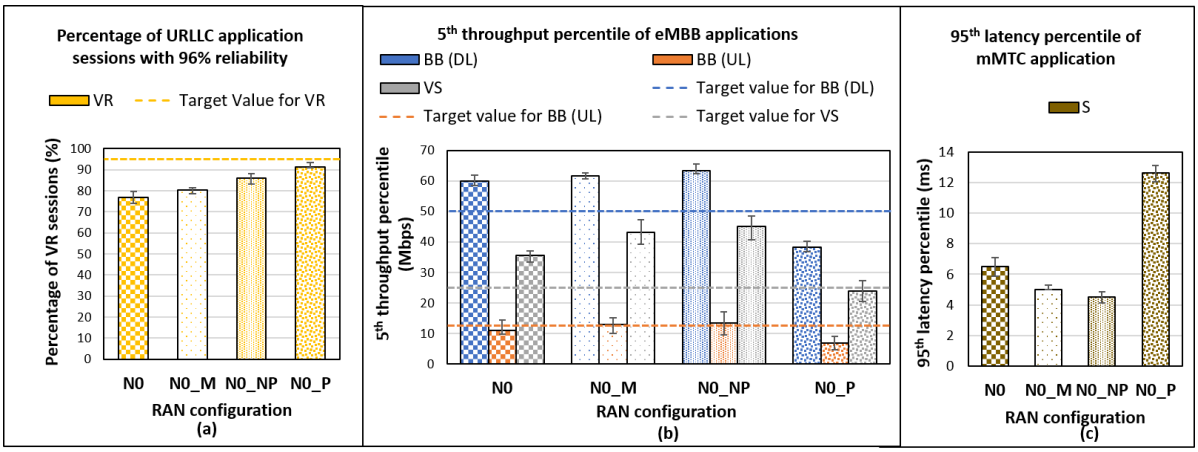

Figure 4.4: KPI values obtained by four RAN configurations, with each configuration using numerology 0 and enabled with regular slot based scheduling or basic mini-slot based scheduling, or non-pre-emptive mini-slot based scheduling or pre-emptive mini-slot based scheduling scheme, respectively.

- *Basic mini-slot based scheduling*: The configuration with basic mini-slot based scheduling scheme is represented as N0_M in Figure 4.4. Figure 4.4.a shows that the performance of the URLLC application (VR) achieved by enabling basic mini-slot based scheduling is better in comparison with the regular slot-based scheduling. This is because mini-slots are shorter scheduling units than the regular slot, and thus allow for faster transmissions as the packet waits for less time before it starts its transmission. Also, eMBB (BB and VS) and mMTC (S) applications benefit from the basic mini-slot based scheduling, as shown in Figure 4.4.b and Figure 4.4.c, respectively. This is because the resources are utilized more efficiently by transmitting multiple packets in distinct mini-slots within a given slot, whereas under regular slot based scheduling, each of these packet transmissions utilize a full slot. Specifically, the use of mini-slots allows for transmission on less than 14 OFDM symbols based on the data size of transmissions, as also explained in details in Section 2.5.2. Also, with the use of mini-slots for transmission, the interference reduces because of lesser symbol usage rather than a full slot of 14 symbols. The target KPI values for only the eMBB (BB and VS) applications are met with the basic mini-slot based scheduling scheme.

- *Non-pre-emptive mini-slot based scheduling*: The configuration with non-pre-emptive mini-slot based scheduling scheme is represented as N0_NP in Figure 4.4. It is observed in Figure 4.4 that for all URLLC (VR), eMBB (BB and VS) and mMTC (S) applications, non-pre-emptive mini-slot based scheduling provides better performance than the basic mini-slot based scheduling. The performance of the URLLC (VR) application improves because the non-pre-emptive mini-slot based scheduling immediately schedules URLLC packets that arrive, in between the regular scheduling moments, if there are idle resources, as explained in Section 2.5.2. Therefore, there is less delay for the URLLC packets in the buffer, which in combination with the potential shortened number of OFDM symbols results in a reduced delay. It is noted that the resource utilization of the URLLC (VR) application increases by 12.4% for non-pre-emptive mini-slot based scheduling in comparison with basic mini-slot based

scheduling, as the effect of scheduling more URLLC packets for achieving less delay in between the regular slots. Considering that the URLLC packets can be transmitted in between the regular slots, there are fewer packets competing for resources at the start of a regular slot, compared to basic mini-slot based scheduling. Therefore, the performance of the eMBB and mMTC applications is also improved. Finally, like with basic mini-slot based scheduling, with non-pre-emptive mini-slot based scheduling the target KPI values are met for only the eMBB applications (BB and VS) but not for the URLLC application (VR).

- *Pre-emptive mini-slot based scheduling*: The configuration with pre-emptive mini-slot based scheduling is represented as N0_P in Figure 4.4. For the URLLC application (VR), this scheduling scheme provides the best performance in comparison with the other mini-slot based scheduling schemes and the regular slot based scheduling, as shown in Figure 4.4.a. This is because the URLLC packets that arrive in between the regular scheduling slots are immediately scheduled, even if there are no idle resources. Specifically, in the case of no idle resources, the scheduler pre-empts an ongoing eMBB or mMTC transmission to transmit the URLLC packets. It is noted that the resource utilization of the URLLC (VR) application is increased by 13.6% in comparison with non-pre-emptive mini-slot based scheduling.

  On the other hand, the pre-emption of the eMBB and the mMTC packets in favor of the URLLC packets, leads to a degrade in the performance of the eMBB (BB and VS) and the mMTC (S) applications because the interrupted eMBB and mMTC packets need to be re-transmitted. It is noted that the total number of retransmissions for the eMBB (BB and VS) and the mMTC (S) applications are increased by 48.4% in comparison with non-pre-emptive mini-slot based scheduling. This results in worse performance for the eMBB and the mMTC applications in comparison with the other mini-slot based scheduling schemes and the regular slot based scheduling, as also shown in Figure 4.4.b and Figure 4.4.c.

  The resulted KPI value obtained for the URLLC (VR) application is the best among the other mini-slot based scheduling schemes and the regular slot based scheduling configurations, but its target value is still not met. For the eMBB (BB and VS) applications, the target KPI value is only achieved for the eMBB (BB (DL)) application.

The pre-emptive mini-slot based scheduling scheme resulted in best among all schemes only for the URLLC application but it is not considered to be used further, because of the performance degradation of the eMBB and mMTC applications by using pre-emptive mini-slot based scheduling. The basic and non-pre-emptive mini-slot based scheduling schemes perform better than the regular slot-based scheduling for all the applications and are thus used further in combination with other features. Specifically, the non-pre-emptive mini-slot based scheduling resulted in yielding best performance among all the schemes, in terms of improving performance for all the applications simultaneously. Therefore, the non-pre-emptive mini-slot based scheduling is selected when the given BWP/slice handles both URLLC and eMBB/mMTC applications, as it schedules the URLLC packets arriving in between the regular slots without any loss to the eMBB and mMTC applications. Technically, the non-pre-emptive mini-slot based scheduling can be used when BWP/slice handle only URLLC applications as well, but with slices and BWPs, it is expected to have lesser resources per BWP/slice and high resource utilization per BWP/slice. The use of non-pre-emptive mini-slot based scheduling will increase the resource utilization and thus result in high interference, which can affect the performance of the applications. Therefore, the choice of configuring the BWP with higher numerology with basic mini-slot based scheduling is considered in further simulations. The basic mini-slot based scheduling is considered, where the URLLC and eMBB/mMTC applications are handled by different BWPs/slices in order to still leverage from the efficient way of allocating resources using mini-slot based scheduling.

In general, in order to achieve low-latency for the URLLC applications we can use mini-slot based scheduling or can use higher numerology. Comparing the performance of these two options, the non-pre-emptive mini-slot based scheduling provides similar performance as the highest numerology configuration (N2). Thus, either of the options can be selected or the options can be combined to further improve the performance of the URLLC application. However, the configuration with numerology 2 and mini-slot based scheduling is not considered in this study because the eMBB and the mMTC applications do not yield good performance with numerology 2 and this study aims to find the overall best configuration for all of the applications, simultaneously. Therefore, the configurations with two BWPs are considered where each BWP can be configured with optimal numerology per application and further combined with mini-slot based scheduling feature.

4.3.1.3. Bandwidth Parts

The previous analysis on configuring the RAN with different numerologies (N0, N1 and N2) concludes that it is difficult to optimally configure the RAN with one suitable numerology for mixed traffic service requirements. In this subsection, the BWP concept, which allows configuring the RAN with multiple numerologies, is assessed. The RAN is divided into multiple BWPs where each BWP is configured with a different numerology than the other BWPs, thus the number of BWPs is determined by the number of different numerologies required based on the service requirements.

Based on the previously obtained results on configuring the RAN with low to high numerology (N0, N1 and N2), shown in Figure 4.3, the optimal numerology for each service category is determined i.e. numerology 0 for the eMBB (BB and VS) and mMTC (S) applications and numerology 2 for the URLLC (VR) application. Thus, the RAN is divided into two BWPs, configured with numerology 0 and numerology 2, respectively, and the division of resources between BWPs is done as explained in Section 3.5.1. Also, a sufficiently large guard band is used in between the two numerologies as explained in Section 2.2 to avoid inter-numerology interference. The URLLC packets are scheduled only on the resources of the BWP configured with numerology 2 and the eMBB and mMTC packets are scheduled only on the resources of the BWP configured with numerology 0. The performance obtained for each application with this configuration is represented as 2B(0,2) in Figure 4.5. Figure 4.5 also shows the previously obtained results with configuration N0 and N2 (the optimal numerology configurations per service group) for comparison purposes.

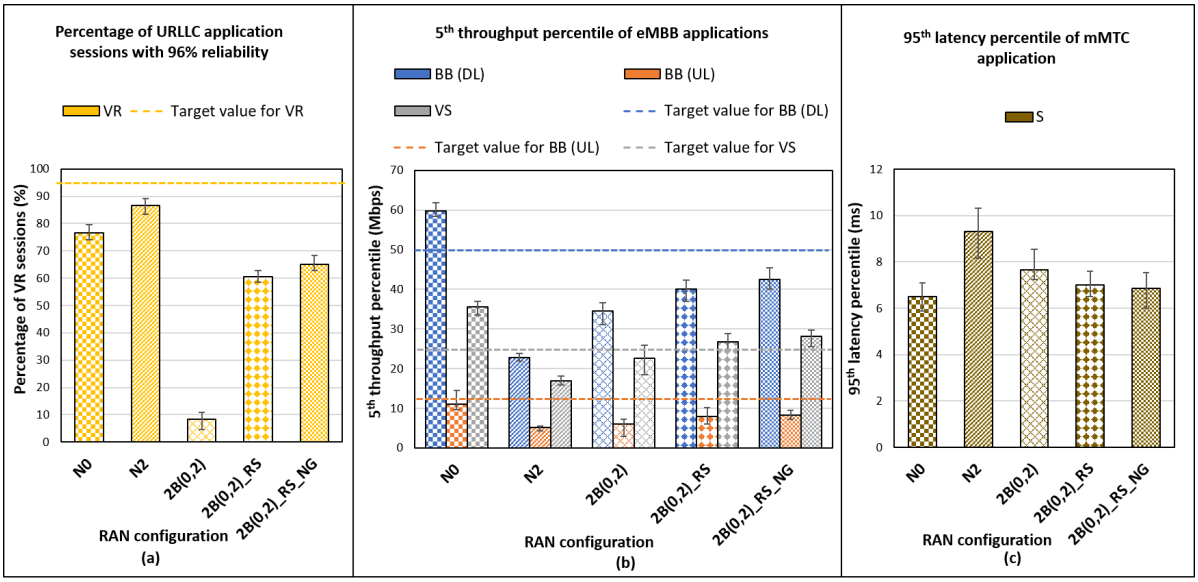

Figure 4.5: KPI values obtained by configurations related to BWPs and comparison with previous configurations with numerology 0 and 2.

Figure 4.5.a shows that the URLLC (VR) application has poor performance when the RAN is divided into two BWPs compared to the undivided RAN configurations. Additionally, the performance of the eMBB (BB and VS) and the mMTC (S) applications is worse with the use of BWPs compared to the undivided RAN configuration with numerology 0 (which is the optimal numerology for these applications), as shown in Figures 4.5.b and 4.5.c. This is because of several reasons: a) The total resources are reduced because of the necessary middle guard band used to avoid INI. b) The reduction in the frequency-selective channel-adaptive scheduling gains because there are fewer resources per BWP. c) Trunking losses due to the split of resources into two BWPs. Because of splitting of resources, the resource utilization is increased and thus, high interference is experienced by the transmission. It is noted that the resource utilization of the BWP with numerology 2 handling the URLLC application is 89.3% and for BWP with numerology 0 handling the eMBB and mMTC applications is 51.2%. This also reflects that the resource split ratio may not be optimal, but optimization of this resource split is outside the scope of this study.

To assess the impact of idle resource sharing between BWPs, the option of allowing resource sharing between the two BWPs (configured with N=0 and N=2), is considered. The results obtained for this configuration are represented as 2B(0,2)_RS in Figure 4.5. Figure 4.5 shows the clear advantage and gains, for all applications, of allowing resource sharing in comparison with the configuration (2B(0,2)) which does not allow for idle resource sharing. This is because the trunking losses due to the split of resources is compensated by enabling the idle resource sharing capability. Also, it is noted that now the resource utilization is balanced between the two BWPs because of enabling the sharing capability. The resource utilization of BWP handling the eMBB and the mMTC applications is 66.1% and of BWP handling URLLC application is 75.4%. Even though the gains of sharing the idle resources are evident, the performance of this configuration (2B(0,2)_RS) is still worse for URLLC (VR) application in comparison with uniformly applied numerology 2 (N2) configuration (the optimal numerology for URLLC application) and for eMBB (BB and VS) and mMTC (S) applications in comparison with uniformly applied numerology 0 (N0) configuration (the optimal numerology for eMBB and mMTC applications).

Next, to investigate the impact of unavailability of resources used for the middle guard band between the two BWPs, on the performance of the applications, simulations are performed without using the middle guard band between the two BWPs, each configured with numerology 0 and 2, respectively and with allowing idle resource sharing. The resulted KPI values, represented as 2B(0,2)_RS_NG in Figure 4.5 show the improvement in the applications' performance in comparison to the configuration with the middle guard band (2B(0,2)_RS). This is because the resources used for the middle guard, in the 2B(0,2)_RS configuration, are now utilized by the applications. Because the guard bands between BWPs are required to avoid INI, in the further considered RAN configurations that use multiple numerologies, this middle guard band is still taken into account.

In an attempt to increase the frequency-selective channel-adaptive scheduling gains and to reduce the resources needed for the middle and edge guard bands, another configuration is considered with a lower numerology for the BWP handling the URLLC application. The RAN is thus divided into two BWPs, configured with numerology 0 and 1, respectively and with resource sharing enabled. Figure 4.6 shows the results obtained by this configuration (2B(0,1)_RS) and it is illustrated that the performance of all the applications is improved in comparison to the configuration 2B(0,2)_RS, where numerology 2 is considered for the BWP handling the URLLC (VR) application.

The improvement observed with configuration 2B(0,1)_RS, and shown in Figure 4.6, in comparison to the configuration 2B(0,2)_RS is because of several reasons: a) The total resources needed for the edge guard band and the middle guard with the configuration 2B(0,1)_RS are 28.37% less in comparison to the configuration 2B(0,2)_RS and therefore more resources are available to schedule data transmissions in both the BWPs. Because of more resources, the frequency-selective channel-adaptive scheduling gains increase with a lower numerology as the total number of PRBs,

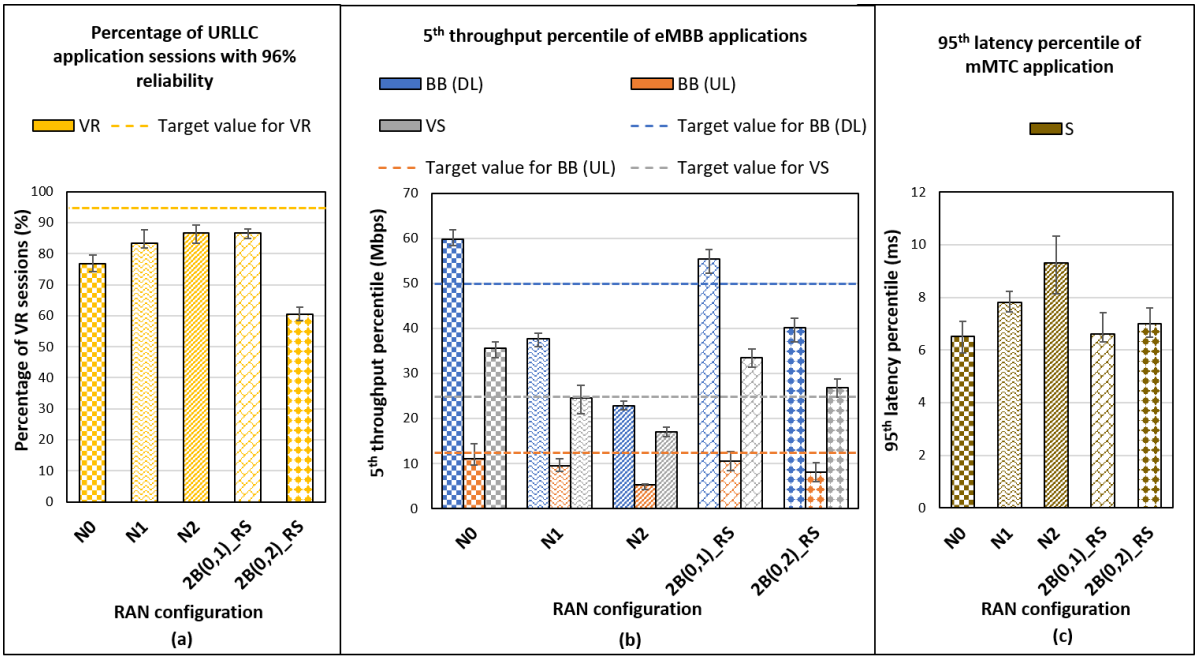

Figure 4.6: KPI values obtained by configurations with BWPs (configured with N=0,1 and N=0,2, respectively) and with the previous configurations for numerology 0,1 and 2.

in the given BWP, is higher. Because of more resources and the higher frequency-selective channel-adaptive scheduling gains, the resource utilization of both BWPs is reduced. It is noted that the resource utilization of the BWP handling the eMBB and mMTC traffic is 62.2% and of the BWP handling the URLLC traffic is 67.6% and thus the transmissions also experience lower interference. The comparison between the configurations 2B(0,1)_RS and 2B(0,2)_RS, reveals that choosing the numerology based on the application type e.g. selecting numerology 2 for the URLLC application, is not necessarily the optimal way because there are other effects of different features which also needs to be taken into account which may eventually lead to better overall performance.

Figure 4.6 also shows that the configuration 2B(0,1)_RS provides better performance for the URLLC (VR) application than uniform configurations with either numerology 0 or 1 and equivalent performance to the configuration with numerology 2 (optimal for the URLLC application for configurations without BWPs). Also, for the eMBB and mMTC applications, except the BB (DL) application, the performance obtained with the configuration 2B(0,1)_RS is almost equivalent to the configuration with numerology 0 (optimal for the eMBB and mMTC applications for configurations without BWPs). This reflects that the BWPs, when configured with appropriate numerologies, and with resource sharing capability between BWPs provide similar performance as to when the RAN was configured with optimal numerologies for URLLC and eMBB/mMTC applications separately. This is because the benefits of both the appropriate numerologies are combined with the use of BWPs and the idle resource sharing between BWPs compensates for the trunking losses due to splitting of resources into BWPs.

Finally, without considering mini-slot based scheduling, the configuration 2B(0,1)_RS is so far the best configuration in terms of handling the requirements of all the applications simultaneously. However, the target KPI values are yet not achieved for the URLLC (VR) and the eMBB (BB (UL)) applications. Also, this configuration 2B(0,1)_RS do not out-perform the optimal numerology configurations N2 and N0 for URLLC and (eMBB and mMTC) applications, respectively.

4.3.1.4. Combination of all possible features

The best so far performing configuration (2B(0,1)_RS) is now combined with the basic mini-slot based scheduling feature to assess the impact of this combination on the performance of the applications. The results of the new configuration (2B(0,1)_RS_M) are shown in Figure 4.7, where it is illustrated that the performance of all the applications is now improved compared to the configuration without mini-slots (2B(0,1)_RS) and to the two configurations with numerology 0 and basic and non-pre-emptive mini-slot based scheduling, respectively.

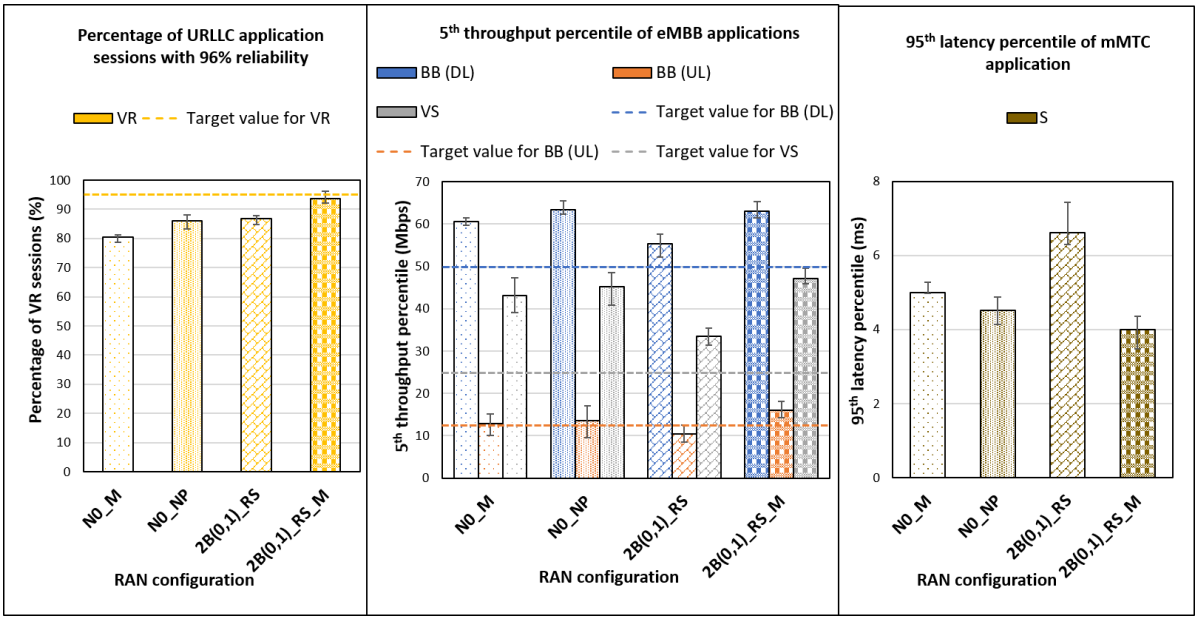

Figure 4.7: KPI values obtained by enabling mini-slot based scheduling schemes with configurations with and without BWPs.

The improvement observed with configuration 2B(0,1)_RS_M in comparison to 2B(0,1)_RS is because the resource allocation is more efficient with mini-slot based scheduling compared to the regular slot based scheduling, as also observed and explained in Section 4.3.1.2. Additionally, the performance of the configuration 2B(0,1)_RS_M is better than the configurations without BWPs (N0_M and N0_NP) because of the benefits that BWPs offer i.e. multiplexing the RAN with the appropriate numerologies, based on the application requirements, with allowing idle resource sharing and then using mini-slots to leverage the benefits of more efficient way of allocating resources by assigning the minimum required symbols for the transmissions.

The last considered configuration (2B(0,1)_RS_M) i.e. combining BWPs, idle resource sharing and mini-slot based scheduling, yields the best performance compared to all the other considered configurations. Specifically, the target KPI values for all the eMBB (BB and VS) applications are achieved and the KPI value of the URLLC (VR) application is very close to its target value.

4.3.2. RAN slices per service category

In this subsection, the RAN slicing concept is considered, where a slice is configured per service category i.e. eMBB, mMTC and URLLC slices as mentioned in Section 2.1. Based on the previous results, the eMBB and mMTC slices are configured with numerology 0 while the URLLC slice is configured with numerology 1 with allowed idle resource sharing between the slices. Also, each RAN slice has a different TDD configuration, based on their DL and UL traffic ratio, which is presented in Section 3.5.2.1. In principle, each slice can also have a different packet scheduler unlike BWPs, as mentioned in Section 2.3, but the effect of it is not seen here, as we considered one packet scheduler for all the slices. The M-LWDF scheduler is applied based on the literature, where it was found suit-

able for mixed-traffic scenarios [24]. The optimization of the scheduler is outside the scope of this study. The mini-slot based scheduling is not yet considered in combination with the three configured slices to make a better and easier comparison between the non-sliced RAN configuration with BWPs and the configuration with three slices. Combining many features together may make unclear which features are responsible for which effects. Also, another similar configuration is considered, but with numerology 2, instead of 1, for the URLLC slice. This is done to confirm the previously made observation for BWPs that numerology 1 is more suitable for the URLLC (VR) application than numerology 2. The results of both configurations are represented as 3S(0,0,1)_RS and 3S(0,0,2)_RS in Figure 4.8, together with the previously obtained results using the BWP configurations 2B(0,1)_RS and 2B(0,2)_RS.

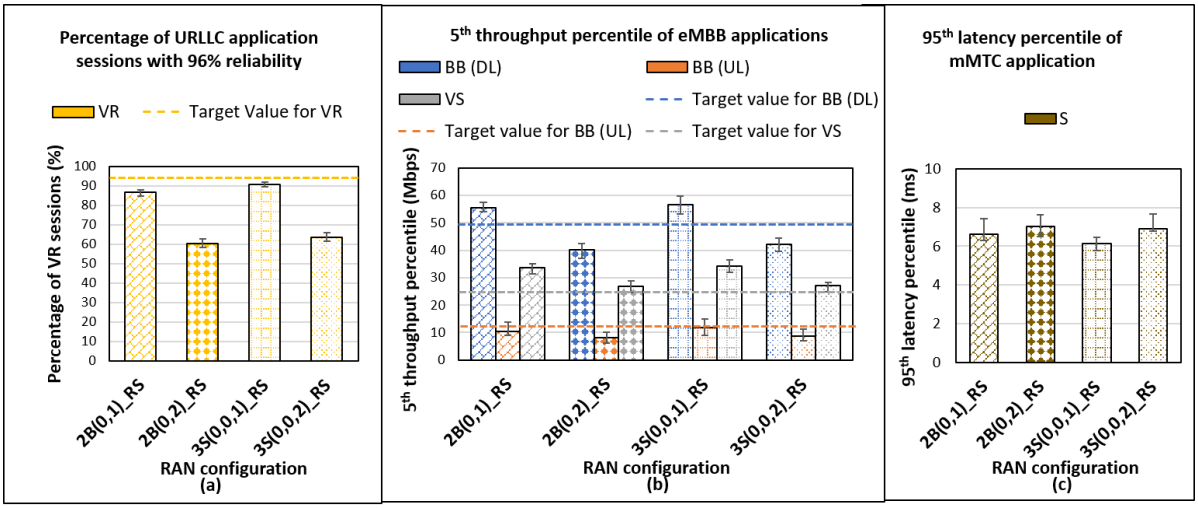

Figure 4.8: KPI values obtained by splitting the RAN into three slices (configured with N=0,0,1 and N=0,0,2) and comparison to the previously considered configurations with BWPs.

Figure 4.8 shows that the configuration 3S(0,0,1)_RS (URLLC slice configured with numerology 1) performs better for all the applications compared to the configuration 3S(0,0,2)_RS (URLLC slice configured with numerology 2). Thus, numerology 1 is considered the best suitable numerology for the URLLC slice in the considered scenario, for the same reasons mentioned in Section 4.3.1.3. Figure 4.8 also shows that both of the RAN slicing configurations (3S(0,0,1)_RS and 3S(0,0,2)_RS) perform slightly better than their respective BWPs configurations (2B(0,1)_RS and 2B(0,2)_RS) for all the applications. This is because each slice is configured with its own TDD configuration based on the slice-specific DL to UL traffic ratio whereas, the BWPs cannot have their own TDD configuration and they thus use the TDD configuration globally determined based on the network DL and UL traffic ratio. Due to this reason, the slices have a more appropriate channel split within a radio frame, for example the mMTC slice will mainly have UL slots, as the traffic is related to the UL channels and thus there are more frequent UL slots for transmissions. The benefit of having different TDD configuration per slice compensates for the drawback of increased trunking losses with splitting the resources into three slices.

With the configuration 3S(0,0,1)_RS, the target KPI values for all the eMBB (BB and VS) applications are met but the target KPI value for the URLLC (VR) application is still not achieved. Therefore, another RAN configuration is considered, which is similar to the 3S(0,0,1)_RS configuration but it additionally uses basic mini-slot based scheduling. Also, in order to reduce the resources used by the middle guard bands between slices, another configuration is considered where all the slices are configured with numerology 0 with allowed resource sharing between slices and basic mini-slot based scheduling is enabled to enhance the performance of the URLLC traffic. The basic mini-slot based scheduling is selected in both configurations because each slice is dedicated to a service cat-

egory and thus none of the slices handle mixed traffic. Using the latter configuration, an evaluation will be made on combining the benefits of additional resources (that were previously used for the guard bands) with the use of mini-slot based scheduling to compensate for a sub-optimal numerology of the URLLC slice.

The resulted KPI values for both of the newly considered configurations are shown in Figure 4.9 (represented as 3S(0,0,1)_RS_M and 3S(0,0,0)_RS_M, respectively) together with the previously obtained results with the configuration 3S(0,0,1)_RS, for comparison purposes. Figure 4.9 shows that the performance of all applications improves with the configuration 3S(0,0,1)_RS_M in comparison to the configuration 3S(0,0,1)_RS, where mini-slot based scheduling is not enabled. This is because of the more efficient resource allocation that basic mini-slot based scheduling offers, as also explained above in Section 4.3.1.2. The results obtained for the configuration 3S(0,0,0)_RS_M show that the additional resources and the use of mini-slot based scheduling is not enough to compensate for using a less suitable numerology for the URLLC slice. Finally, Figure 4.9 illustrates that *the RAN slicing configuration 3S(0,0,1)_RS_M is the first configuration so far with which the target KPI values for all eMBB (BB and VS) and URLLC (VR) applications are achieved.* Thus, this configuration is considered the best performing configuration thus far.

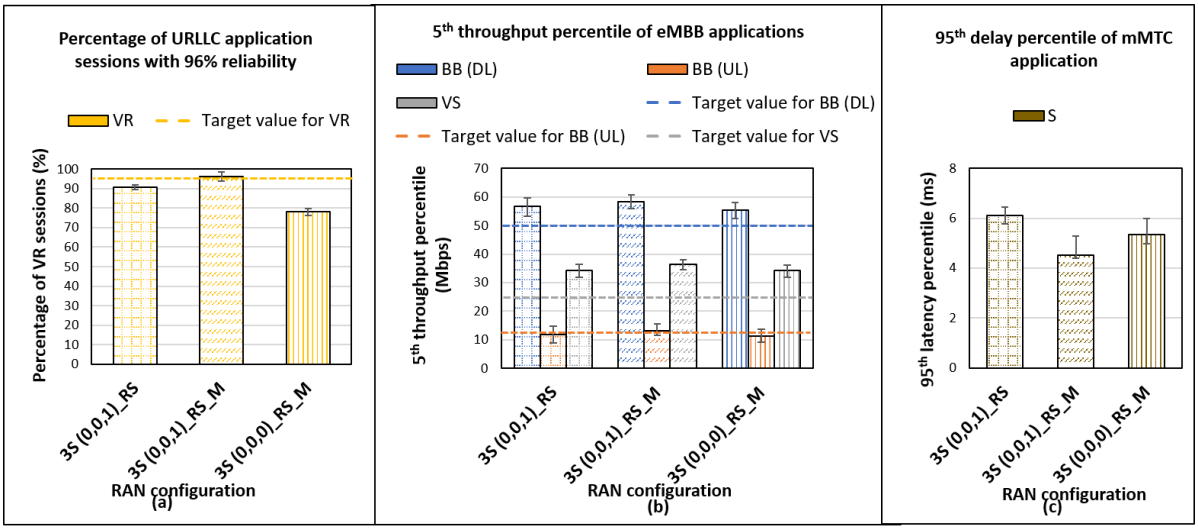

Figure 4.9: KPI values obtained by splitting the RAN into three slices (configured with N=0,0,0 and N=0,0,1) with the latter configured with and without mini-slot based scheduling.

4.3.3. RAN slices per service category and for a customer

This study also aims to analyze the benefits and/or losses of configuring a separate slice for the emergency service group. The applications associated with the emergency service group, when there are no ongoing incidents, are VS and sensors sending risk-sensitive measurements. The analysis is done by investigating a configuration that consists of four slices: three slices, one per service category i.e. URLLC, eMBB and mMTC and a separate slice for the emergency service group, also called the customer slice. Also, the resource sharing capability between the slices and the basic mini-slot based scheduling are enabled. Additionally, each RAN slice has a different TDD configuration, based on its DL and UL traffic ratio, which is presented in Section 3.5.2.2. Similar to other slicing configurations, M-LWDF scheduler is used for all the slices. The RAN slices dedicated to the service categories are configured with the best suitable numerologies as in configuration 3S(0,0,1)_RS, presented in Section 4.3.2, i.e. the eMBB and mMTC slices are configured with numerology 0 and the URLLC slice is configured with numerology 1. For the slice dedicated to the emergency service group, numerology 0 is considered as the suitable numerology, as the slice consists of only eMBB

(VS) and mMTC (S) traffic. The results obtained with this configuration are shown in Figure 4.10, represented as 4S(0,0,0,1)_RS_M, next to the previously obtained results with the BWPs and RAN slicing configuration with three slices, 2B(0,1)_RS_M) and 3S(0,0,1)_RS_M, respectively to compare the three architectural choices.

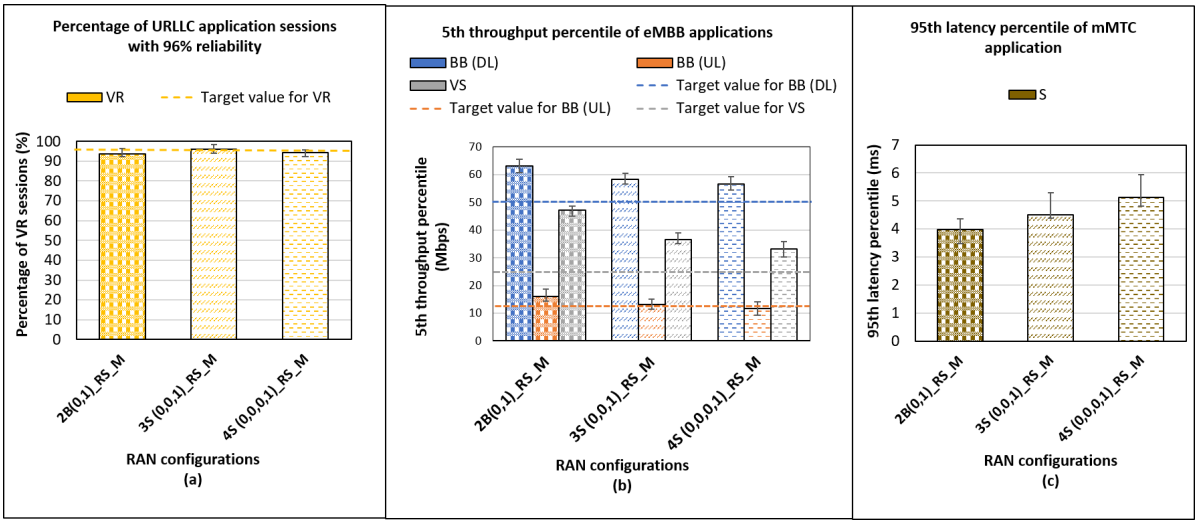

Figure 4.10: KPI values obtained by splitting the RAN into three slices (configured with N=0,0,1) and four slices (configured with N=0,0,0,1 and) with resource sharing and mini-slot based scheduling enabled for both configurations.

Figure 4.10 shows that the performance of all the applications under the configuration with four slices (4S(0,0,0,1)_RS_M) is slightly worse than the performance under the configuration with three slices (3S(0,0,1)_RS_M). This is because of the trunking losses of further splitting the resources, because of the additional customer slice. Due to this split of resources, the assigned number of PRBs to eMBB and mMTC slice is smaller in comparison with the configuration with three slices. Therefore, the frequency-selective channel-adaptive scheduling gains are also reduced. The target KPI values are still met for all the applications taking into account the confidence interval, specifically for the BB (UL) application.

Comparing the results of BWPs and slicing configurations, the BWPs configuration (2B(0,1)_RS_M)) performs better for the eMBB and mMTC applications whereas, the slicing configurations ((3S(0,0,1)_RS_M) and (4S(0,0,0,1)_RS_M)) perform better for the URLLC application. This is because the M-LWDF scheduler follow a weighted version of PF scheduler for URLLC traffic because of which, the scheduler schedules the URLLC traffic first, if given a choice between URLLC/eMBB/mMTC traffic with same achievable bit rate in a given TTI, while sharing the resources between slices. The benefit for the URLLC traffic by the scheduler is evidently seen with the use of mini-slots because the resources are assigned more efficiently and there are more sharing opportunity with use of mini-slots based scheduling. The slicing offers similar performance like BWPs but here, the slicing configurations yields better performance as all of the applications achieve the target value.

4.4. Assessment of the impact of an incident in the network

This study also aims to assess the impact of the incident on the performance of the applications and to find the best RAN configuration for the incident scenario. It is reminded that AR, SE and BC are the incident-based applications. To find the impact of an incident, the configurations which performed the best among the non-sliced and the two slicing options, are considered for this assessment. Additionally, some more configurations are considered, specifically with the four slice option based on the mixed application requirements for the customer slice, which are explained further in

this section.

The incident occurs in only one of the 57 cells, which is why the results for the incident scenario are presented based on the performance of the incident cell. Also, the results for the incident scenario are represented by adding the notation in parentheses '(I)' to the configurations to distinguish the results obtained for normal and incident scenario.

4.4.1. Non-sliced RAN configurations

The simulations are performed using the best performing non-sliced RAN configuration (2B(0,1)_RS_M) for the normal scenario, which is the configuration where the RAN is configured with two BWPs, one with numerology 0 and the other one with numerology 1, and with idle resource sharing and basic mini-slot based scheduling enabled. However, the TDD configuration used for the normal scenario is different than for the incident scenario because of the change in the traffic load, as presented in Section 3.5.1. Also, another RAN configuration is considered with the same set of features as in 2B(0,1)_RS_M, with the difference that numerology 2 is used in the BWP handling the URLLC traffic, to find the more appropriate numerology for the incident-based URLLC (AR and SE) applications. The performance of the applications in the incident cell for both of the configurations are presented as 2B(0,1)_RS_M (I) and 2B(0,2)_RS_M (I) in Figure 4.11, next to the previously obtained results of the configuration 2B(0,1)_RS_M, for the normal scenario.

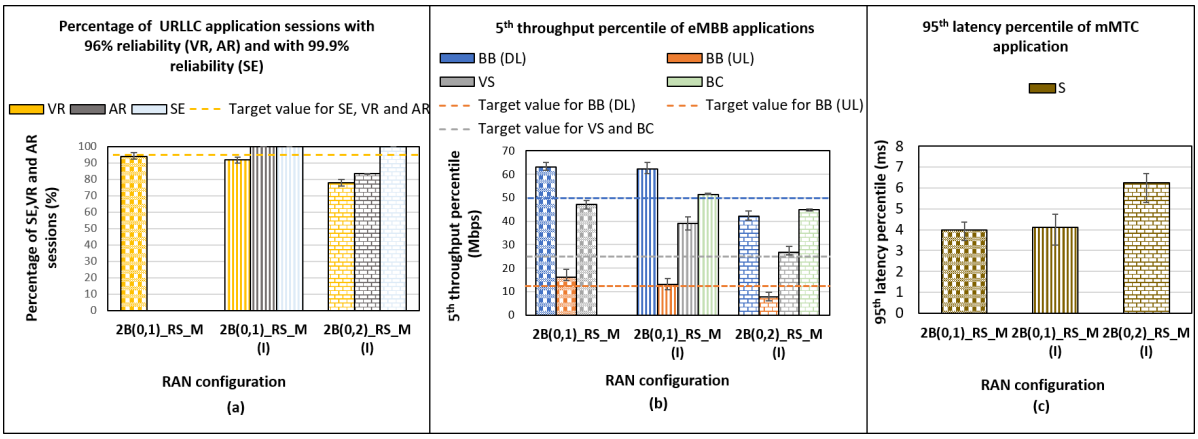

Figure 4.11: KPI values obtained by non-sliced RAN configurations for normal and incident scenarios.

Figure 4.11 shows that the performance of the non-incident related applications under both configurations considered for the incident scenario worsen when an incident occurs (compared to not having an incident). This is because the incident introduces new traffic to the incident-cell because of which there is more competition for resources in the incident cell for all the applications. It is also noted that the resource utilization for configuration 2B(0,1)_RS_M (I) with incident is 74.9% and 83.2% which is 12.4% and 14.1% higher for BWP-1 and BWP-2, respectively, than the same configuration without incident because of increase traffic load.

Figure 4.11 also shows that the performance under the configuration 2B(0,2)_RS_M (I) (with numerology 2 for the URLLC traffic) for all the applications except SE worsens compared to the configuration 2B(0,1)_RS_M (I) (with numerology 1 for URLLC traffic). Thus, numerology 1 is the suitable numerology for the incident-based applications with BWPs configuration and will also be considered with the slicing configurations. Also, the target KPI values for all the applications except for the URLLC (VR) application are achieved under the 2B(0,1)_RS_M (I) configuration.

4.4.2. RAN slices per service category

The simulations are performed using the found best-performing RAN configuration with three RAN slices, without an incident, where eMBB and mMTC slices are configured with numerology 0 and URLLC slice configured with numerology 1 with enabled resource sharing and mini-slot based scheduling together with the incident. The TDD configuration for this configuration is presented in Section 3.5.2.1.

The resulted KPI values are represented as 3S(0,0,1)_RS_M (I) in Figure 4.12 together with the same configuration without an incident (3S(0,0,1)_RS_M) to see the impact of an incident. Figure 4.12 shows that the performance of all the non-incident related applications decreases in comparison with the configuration 3S(0,0,1)_RS_M which did not consider an incident. This drop in performance is because of the increased traffic load due to incident-based applications, which increases the competition for resources in the incident-cell for all the applications, as also explained in above Section 4.4.1. The target KPI values for all the applications are met except for the eMBB (BB (UL)) application.

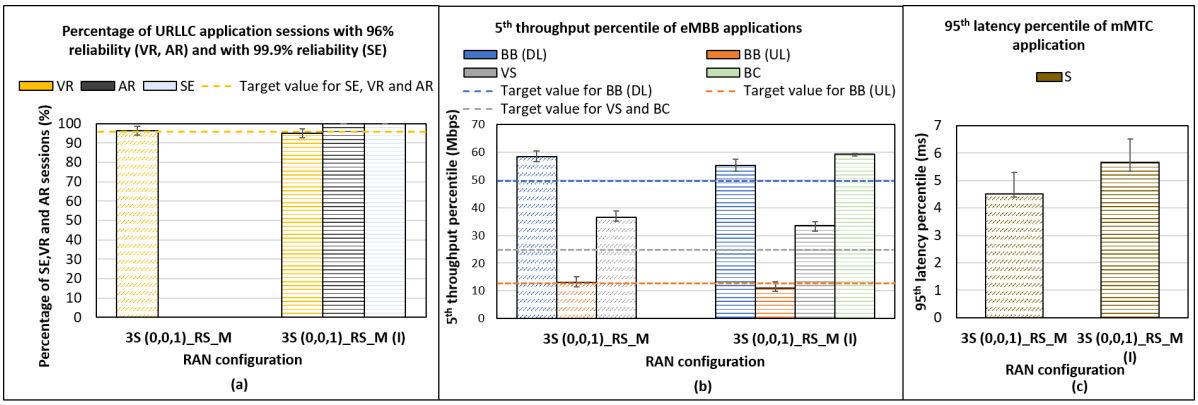

Figure 4.12: KPI values obtained by RAN slice configurations with three slices for normal and incident scenarios.

4.4.3. RAN slices per service category and for a customer

The incident-based applications for this slicing choice are handled by the slice dedicated to a customer, which in this study is the emergency service group. The customer slice was configured with numerology 0 for the scenario without an incident because of only handling eMBB and mMTC traffic, as mentioned in subsection 4.3.3. Now, for the scenario with an incident, the customer slice has traffic with mixed requirements i.e. URLLC traffic (AR and SE applications), eMBB traffic (VS and BC application) and mMTC traffic (sensors sending risk-sensitive measurements). Therefore, to find the most suitable numerology for this slice, three configurations are considered, one with each of the three possible numerologies. The other three slices are configured with the previously found (see Section 4.3.2) optimal numerologies, i.e. numerology 0 for the eMBB and the mMTC slices and numerology 1 for the URLLC slice. Additionally, the idle resource sharing capability between the slices is enabled and the TDD configuration is adjusted based on the DL to UL traffic ratio of each slice, as presented in Section 3.5.2.2. Figure 4.13 shows the results obtained for the three RAN configurations, represented as 4S(0,0,0,1)_RS (I), 4S(0,0,1,1)_RS (I) and 4S(0,0,2,1)_RS (I).

It is observed from Figure 4.13 that the URLLC (AR and SE) applications performs best when the customer slice is configured with numerology 1 but the eMBB (BC) application performs best, when the customer slice is configured with numerology 0. The URLLC applications perform better with numerology 1 in comparison with numerology 2 because in configuration 4S(0,0,1,1)_RS (I), only one middle guard is used between numerology 0 and 1 sub-carriers whereas in configuration 4S(0,0,2,1)_RS (I), two middle guard bands are used between numerologies 0 and 2 and numerolo-

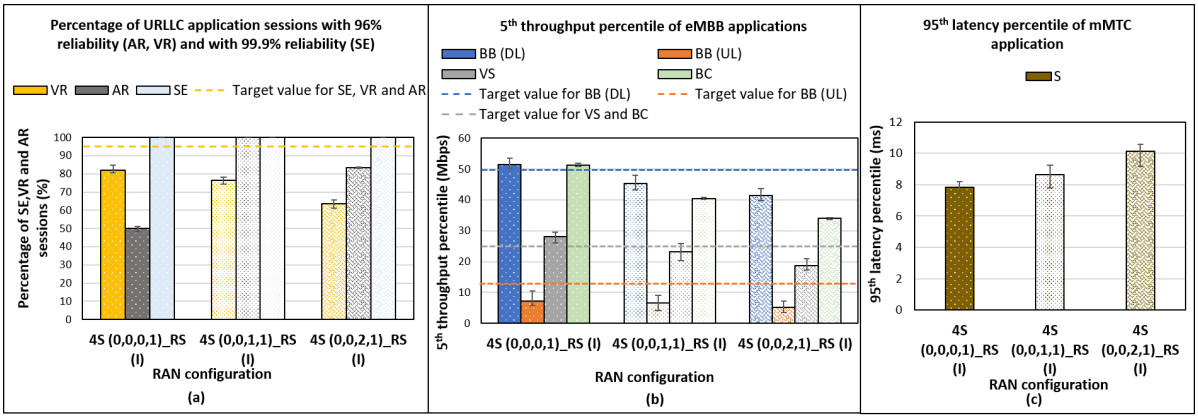

Figure 4.13: KPI values obtained by configurations with four slices, where the RAN is configured with N=0,0,0,1, N=0,0,1,1 and N=0,0,2,1, respectively, for incident scenario.

gies 1 and 2. Therefore, more resources are wasted, which resulted in worse performance with numerology 2. It is concluded that there is not one optimal numerology found for the customer slice, as expected due to the mixed application requirements.

To achieve a better performance for the customer slice, there are two options that can be considered: a) The non-pre-emptive mini-slot based scheduling can be enabled to improve the performance of URLLC (AR) application with configuration 4S(0,0,0,1)_RS (I). b) The customer slice can be split into two BWPs, where each BWP can be configured with appropriate numerologies based on the customer slice application requirements. Both of these options are considered.

Considering the first option, the configuration 4S(0,0,0,1)_RS (I) where the eMBB, mMTC, customer and URLLC slices are configured with numerology 0,0,0 and 1, respectively with allowed idle resource sharing capability is combined with non-pre-emptive mini-slot based scheduling. The resulted KPI values are represented as 4S(0,0,0,1)_RS_NP (I) in Figure 4.14 shows that the performance of URLLC (AR) application is improved in comparison with configuration 4S(0,0,0,1)_RS (I), where regular slot based scheduling is enabled. This is because the URLLC packets are scheduled immediately as they arrive in the buffer in between the regular scheduling slots, as also explained in Section 4.3.1.2.

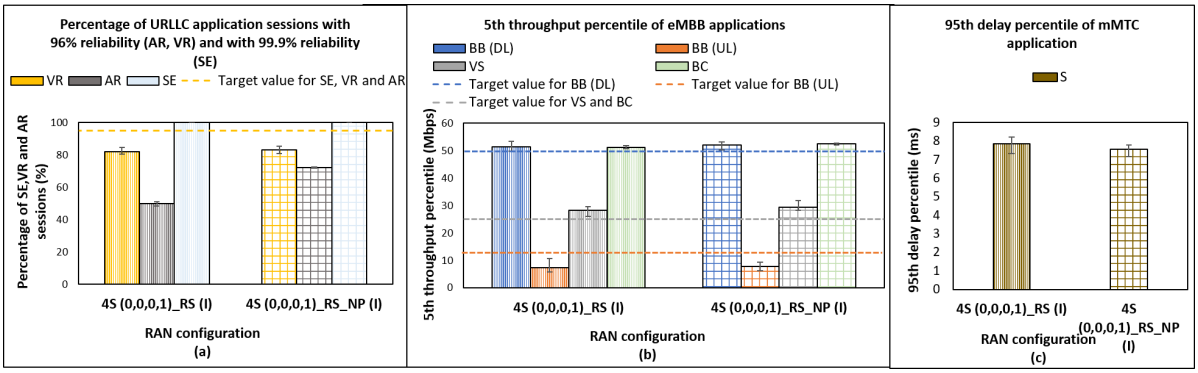

Figure 4.14: KPI values obtained by configurations with four RAN slices, configured with N=0,0,0,1 with and without non-pre-emptive mini-slot based scheduling for customer slice.

Now, considering the second option, we consider a new configuration where the customer slice is split into two BWPs: one BWP for handling the URLLC traffic configured with numerology 1 and the other BWP for handling the eMBB and the mMTC traffic configured with numerology 0. Similar to before, the resource sharing capability is enabled and the TDD configuration remains the same.

The results obtained with this configuration are represented as 4S(0,0,2B(0,1),1)_RS (I) in Figure 4.15. To analyze the benefits of splitting the customer slice into BWPs, the previously obtained results of the configuration 4S(0,0,0,1)_RS_NP (I) are also shown in Figure 4.15 to make a comparison between the two options of improving the performance of customer slice applications.

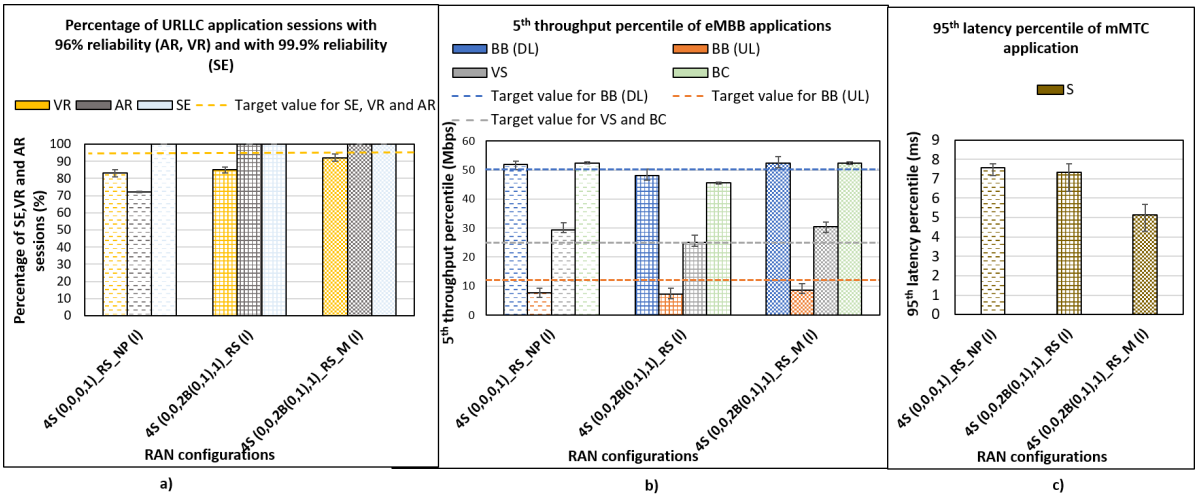

Figure 4.15: KPI values obtained by configurations with four RAN slices, configured with N=0,0,0,1 and N=0,0,(0,1),1, respectively with former enabled with non-pre-emptive mini-slot based scheduling and latter enabled with and without mini-slot based scheduling.

Figure 4.15 shows that the configuration where the customer slice is split into two BWPs (4S(0,0,2B(0,1),1)_RS (I)), with numerologies 0 and 1, performs better for URLLC (AR) application and similar for URLLC (SE) application for customer slice in comparison with configuration, where customer slice is configured with one numerology 0 with non-pre-emptive mini-slot based scheduling 4S(0,0,0,1)_RS_NP (I). The applications of the customer slice perform better because the resources assigned to the customer slice is split into two BWPs which are configured with a suitable numerology. The splitting of the slice's resources brings two losses, it reduces the frequency-selective channel adaptive scheduling and trunking gains in comparison with the case where the customer slice is configured with one numerology. However, the gains obtained by multiplexing two appropriate numerologies and sharing the idle resources between the BWPs are more than the two losses. The other applications perform similar in comparison between the two configurations 4S(0,0,0,1)_RS_NP (I) and (4S(0,0,2B(0,1),1)_RS (I)) and thus, it is noted that multiplexing appropriate numerologies, within the given slice, is more beneficial than configuring the RAN with one numerology and using non-pre-emptive mini-slot based scheduling.

Another configuration is considered which is the same as configuration 4S(0,0,2B(0,1),1)_RS (I) with the addition of enabling mini-slot based scheduling and the results from this configuration are represented as 4S(0,0,2B(0,1),1)_RS_M (I) in Figure 4.15. Figure 4.15 shows the improvement in performance for all the applications for the configuration which uses mini-slot based scheduling (4S(0,0,2B(0,1),1)_RS_M (I)) in comparison to the configuration (4S(0,0,2B(0,1),1)_RS (I)) which uses regular slot based scheduling. This is expected because of the benefits obtained by assigning resources more efficiently with mini-slot based scheduling, as explained in Section 2.5.2.

The KPI values achieved by the four slice RAN configuration 4S(0,0,2B(0,1),1)_RS_M (I), i.e. the customer slice is split into two BWPs and resource sharing and mini-slot based scheduling is enabled are better than all the other considered four slice RAN configurations, for the scenario with an incident. However, the target KPI value for URLLC (VR) and eMBB (BB (UL)) applications are not met.

Figure 4.16 shows the results obtained by best-performing configurations under each architec-

tural choice for the incident scenario. The results show that the URLLC applications perform better with slicing configuration than the configuration with BWPs whereas, the eMBB and mMTC applications perform better with the configuration with BWPs than slicing configurations. This is because of the same reason, as explained with the normal scenario case. In comparison with all three cases, none of the configuration achieve the target KPI value of all the applications simultaneously.

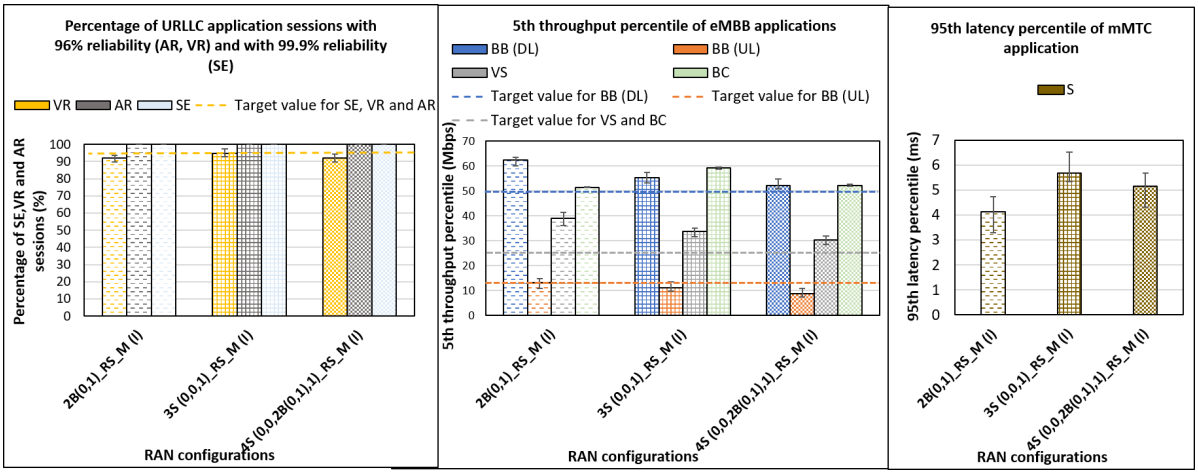

Figure 4.16: KPI values obtained by three configurations with two BWPs, three slices and four slices, respectively for incident scenario.

4.5. Best-performing RAN configurations for normal and incident scenarios

In the previous sections, different RAN configurations have been investigated to serve mixed traffic based on the normal and incident scenario. All the RAN configurations are visualized together for each application and the best-performing configurations are determined in this section for both scenarios. The configuration with which all the target KPI values are achieved for all the applications is determined as the best-performing configuration. In case, multiple RAN configurations achieves all the target KPI values of each application, the configuration which yields better performance among those configurations are considered the best-performing configuration.

Figures 4.17-4.22 illustrate an aggregation of all the previously obtained performance by all the considered RAN configurations for both normal and incident scenarios, for all the applications. The RAN configuration 3S(0,0,1)_RS_M, which consists of three slices, one per service category, with each slice configured with the best suitable numerology i.e. 0 for the eMBB and mMTC slices and 1 for the URLLC slice, and with enabled mini-slot based scheduling and idle resource sharing between slices, is the only configuration which achieves the target KPI values for all the applications simultaneously, just by considering the mean values. The degree of certainty of achieving the target KPIs for all the applications is highest with this configuration among all the considered configurations. Thus, this RAN configuration is considered the best among all the RAN configurations defined and evaluated in this study for the normal scenario.

For the scenario with an incident, none of the considered RAN configurations are able to meet all the target KPI values for all the applications simultaneously. Therefore, the RAN configurations which are able to achieve all the applications except for only one application is then considered and there are two such configurations, which are: a) The RAN configuration (2B(0,1)_RS_M (I)) with two BWPs, each configured with numerology 0 and 1 to handle (eMBB, mMTC) and URLLC traffic, respectively with enabled mini-slot based scheduling and idle resource sharing capabilities which achieves the target performance of all the non-incident and incident-based applications ex-

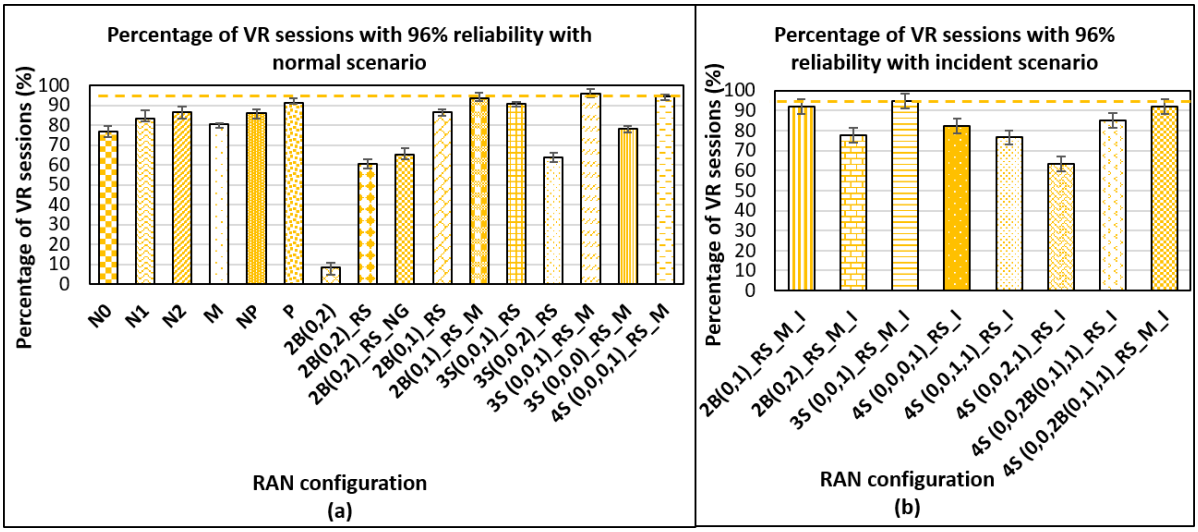

Figure 4.17: KPI values obtained for VR sessions by all RAN configurations simulated in this study.

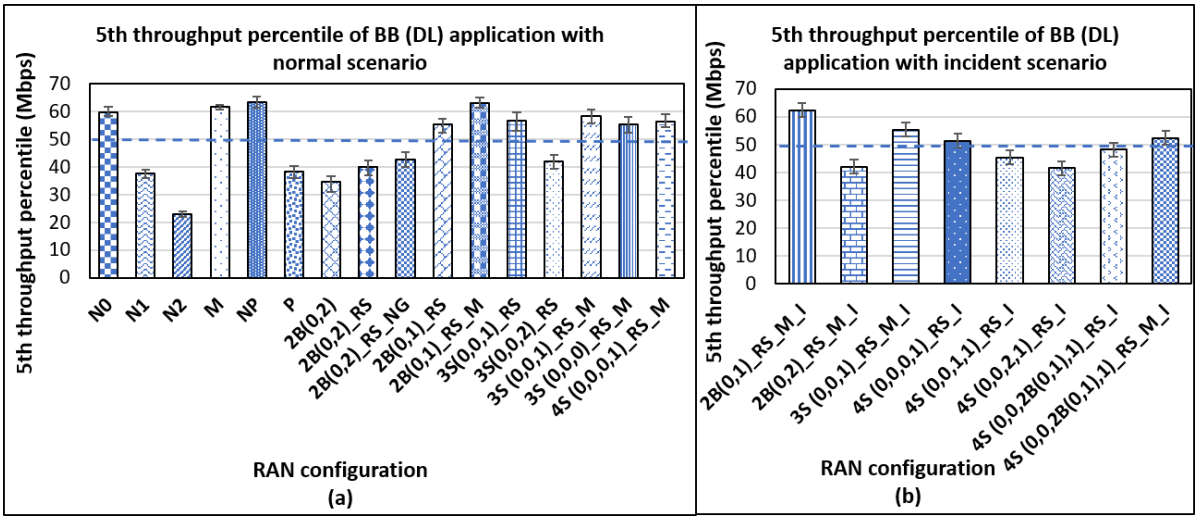

Figure 4.18: KPI values obtained for broadband access (DL) application by all RAN configurations simulated in this study.

cept for URLLC (VR) application. b) The RAN configuration (3S(0,0,1)_RS_M (I)) with three slices, where eMBB and mMTC slices are configured with numerology 0 and URLLC slice configured with numerology 1 with enabled resource sharing capabilities and mini-slot based scheduling, which achieves the target KPI values for all the non-incident and incident-based applications except for the eMBB (BB (UL)) application. The configuration 3S(0,0,1)_RS_M (I) has the same architecture and features as the best performing RAN configuration for normal scenario without an incident. This reassures it is among the best configuration for the incident scenario. Although, between the configuration (2B(0,1)_RS_M (I)) and 3S(0,0,1)_RS_M (I), the respective applications which did not achieve the target which are VR and BB (UL), the VR application is more closer to the target value. Therefore, only the incident-cell can be reconfigured with the configuration (2B(0,1)_RS_M (I)) when an incident occurs, keeping the same architecture and same features as best-performing configuration 3S(0,0,1)_RS_M with normal scenario for all other cells in the network.

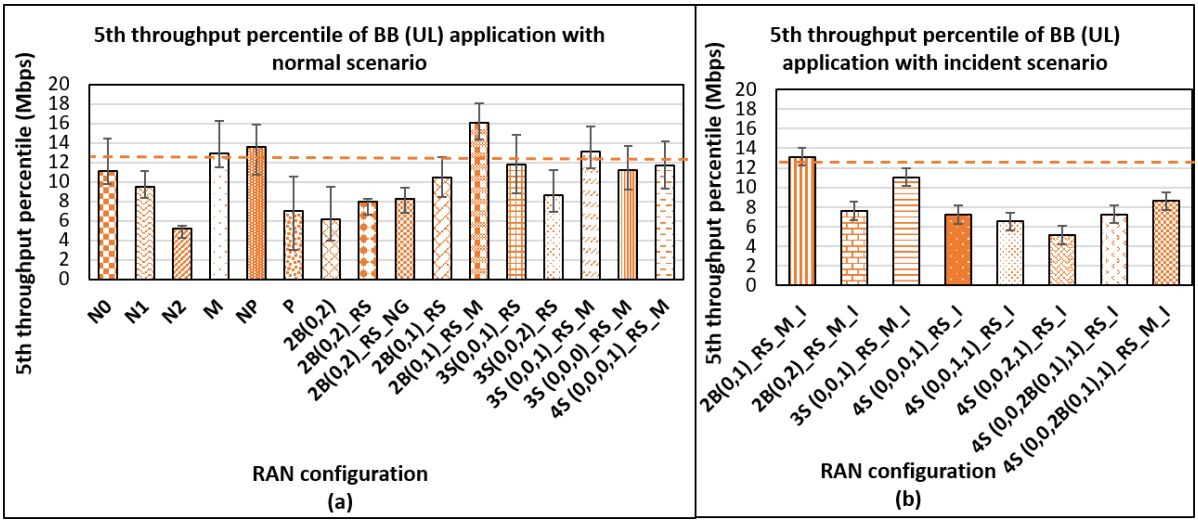

Figure 4.19: KPI values obtained for broadband access (UL) application by all RAN configurations simulated in this study.

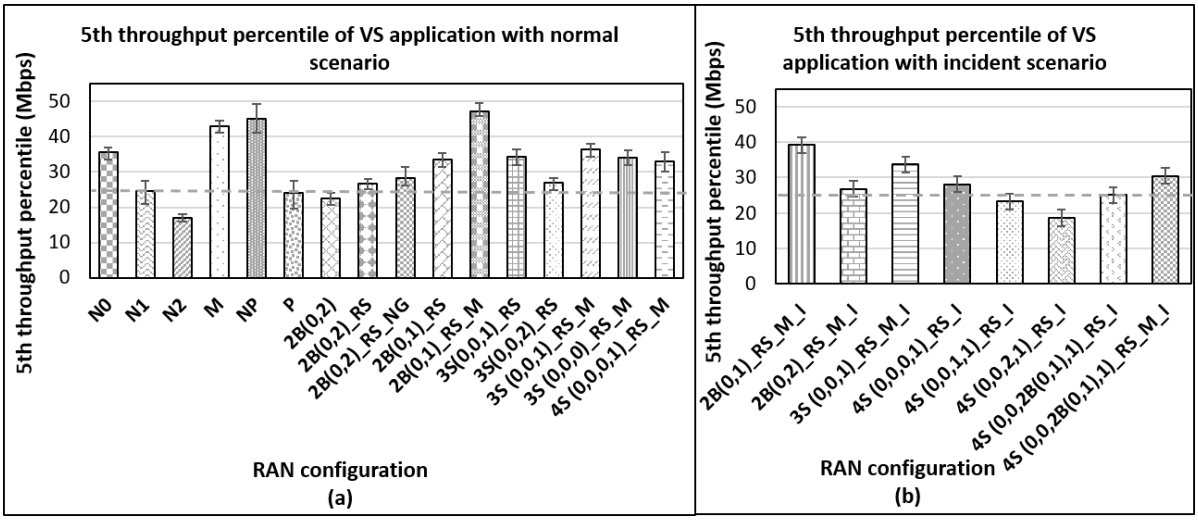

Figure 4.20: KPI values obtained for VS application by all RAN configurations simulated in this study.

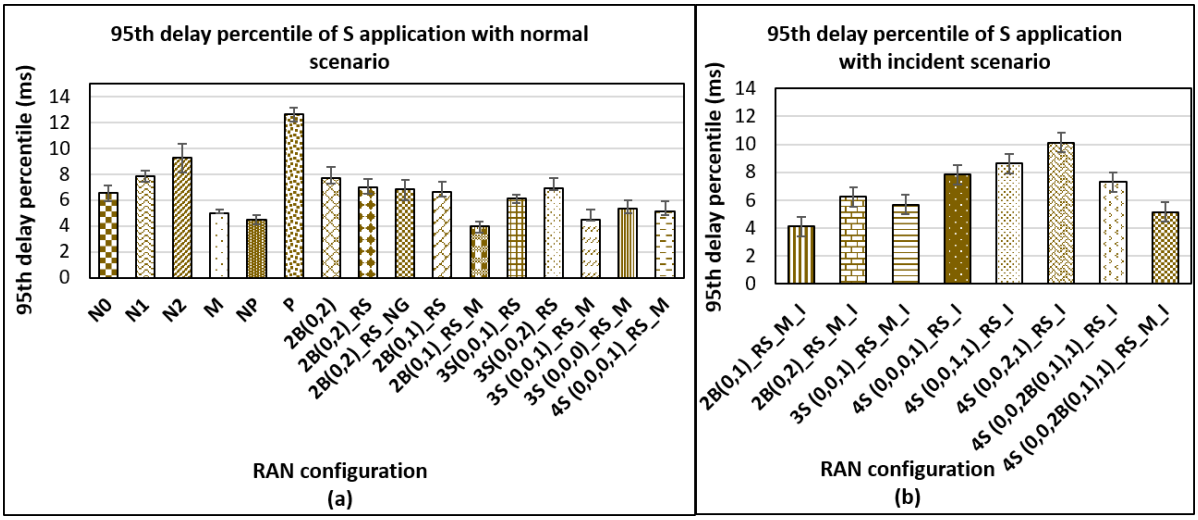

Figure 4.21: KPI values obtained for sensors by all RAN configurations simulated in this study.

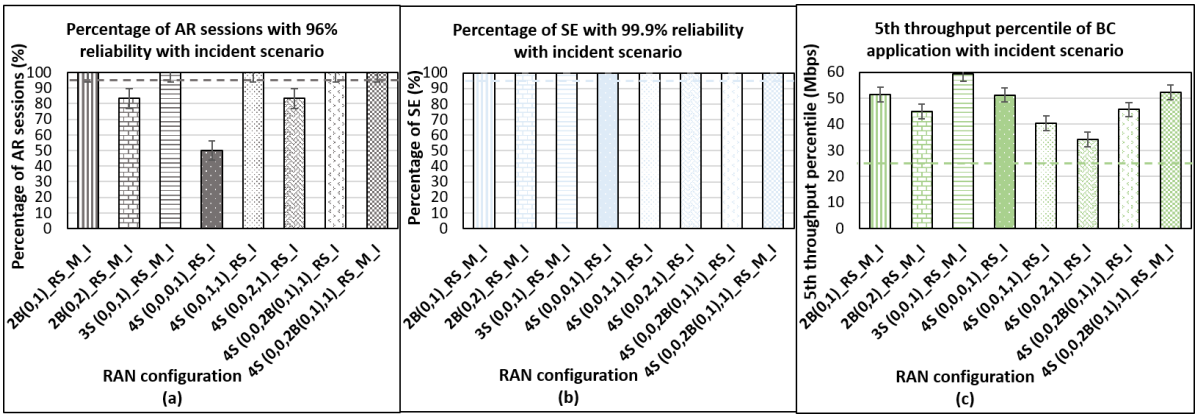

Figure 4.22: KPI values obtained for incident-based applications by all RAN configurations simulated in this study.

5 Concluding Remarks

This chapter provides the summary of this study and highlights the main findings of the research. The main conclusions based on the conducted analysis are presented in Section 5.1 and the recommendations for the future work are provided in Section 5.2.

This study addresses the complex problem of managing and supporting the three main service categories namely eMBB, mMTC and URLLC in a 5G network simultaneously, by assessing different 5G RAN features for resource provisioning, in a smart city urban macro-cellular environment. The evaluation was carried out by configuring the RAN with different sets of RAN features and then comparing the performance of the considered applications for each of these RAN configurations. Also, scenarios with and without an incident were considered to compare the impact of an incident on the network.

5.1. Conclusions

Based on the evaluation of all the individual or combinations of the RAN features, the main conclusions on merits of these RAN features are made. The key findings about each RAN feature are listed below:

- **Flexible numerology:** Configuring the RAN with a higher numerology is beneficial for the URLLC applications because of the reduced slot duration with an increase in numerology, as it is seen that 12.9 % more devices with URLLC (VR) application achieves the reliability targets with numerology 2 in comparison with numerology 0 in case of uniformly applied numerology in the whole carrier. Whereas, a lower numerology is beneficial for the eMBB and the mMTC applications because of having more PRBs, in the given bandwidth, which yields more frequency-selective channel adaptive scheduling gains. The trade-off between latency and throughput is confirmed to be true. Therefore, configuring the RAN with a tailored numerology per service category when there are applications of all the three service categories enhances good performance.

- **Mini-slot based scheduling:** The mini-slot based scheduling is used to achieve the low-latency performance requirement of the URLLC applications. The three considered mini-slot based scheduling schemes yield better performance than regular slot based scheduling for the URLLC application because of having shorter delay until transmission, shorter actual transmissions with using lesser symbols, more efficient use of resources and reduced interference. Also, the basic and non-pre-emptive mini-slot based scheduling schemes are also advantageous for the eMBB and the mMTC applications as the resource assignment is done more efficiently than regular slot based scheduling. Regarding pre-emptive mini-slot based scheduling, a degradation on the performance of the eMBB and the mMTC applications is observed because of the required retransmissions of the pre-empted eMBB and mMTC packets interrupted to serve and prioritize the URLLC application packets. The non-pre-emptive mini-slot based scheduling is concluded to be best for achieving the target performance of all the three service categories, simultaneously.

- **Bandwidth Parts:** The concept of BWPs is beneficial in multiplexing the RAN with different numerologies but only if idle resource sharing between BWPs is allowed. Otherwise, the experienced trunking and frequency-selective channel-adaptive scheduling losses are too significant. The choice of the numerology per BWP should not be considered blindly based on the

optimal numerology per application requirement. The effects of the BWP feature should be taken into account, such as the amount of resources required by the middle and edge guard bands for the considered numerologies and the above-mentioned two losses due to splitting resources between the BWPs. The performance with BWP is similar to slicing, the BWP concept provides the flexibility of customizing the non-sliced RAN or a slice based on different service requirements in the network or in a slice.

- **RAN slicing:** Slicing the RAN is beneficial in managing and achieving the performance requirements of each service category. Among the two considered RAN slicing architecture, viz. applying a RAN slice per service category and applying a RAN slice per service category and a separate slice for customer, the former option is more beneficial and it is concluded that there are more trunking and frequency-selective channel adaptive scheduling losses with an increase in the number of slices because of the additional split of resources. Also, like BWPs, slicing the RAN is only beneficial if idle resource sharing is allowed between slices. The same observation with the choice of numerology is made, like BWPs, that the selection of numerology per slice should consider the effects of guard bands and the trunking and scheduling losses. It is also noted that slicing offers additional improvement in achieving the target performance with the flexibility of having different TDD configuration per slice. Although, in operational networks, the regulatory enforces a specific split to be used by all the operators, as also mentioned in Section 2.4 which removes the flexibility of tailoring the TDD split to the estimated traffic symmetry. Also, in this study only one packet scheduler is considered but potentially, slicing offers the benefit of selecting different packet schedulers as per the slice-specific requirement.

- **Inter-slice/BWPs resource sharing:** The idle resource sharing between slices and/or BWPs is concluded to be essential in achieving the target performance. This is because it compensates for the trunking losses which otherwise occur due to the split of resources between the slices and/or BWPs.

Further, there are some key findings obtained by the different considered RAN configurations, where each configuration consists of a combination of features. Firstly, combining different features is beneficial as one feature can compensate for the losses of another feature and thus helps in achieving the target performance, for example, by configuring the RAN with a lower numerology which fits best to achieve the requirements of the eMBB and the mMTC traffic and enabling mini-slot based scheduling for achieving the low-latency requirements of the URLLC traffic. Secondly, it is concluded that configuring the RAN with BWPs and combining with mini-slot based scheduling and idle resource sharing capability is more beneficial than configuring the RAN with a single numerology and mini-slot based scheduling. Lastly, from the comparison between RAN slicing and BWPs, it is concluded that RAN slicing yields almost similar performance as with the use of BWPs but RAN slicing provides ease of managing different service category requirements on the same physical infrastructure. For example, some tenants or customers require high security with the transmission and with slicing concept, the operator can visibly provide more security to individual service class or tenants.

The study investigated the normal scenario and the scenario with an incident. Among all the considered RAN configurations, the best-performing configurations were determined for both the normal and the incident scenario. For the normal scenario, the configuration where the RAN was configured with three slices, one per service category, in combination with enabled mini-slot based scheduling and with allowed idle resource sharing capability resulted in the best overall performance. For the incident scenario, two configurations, one with same architecture and feature as the best-performing configuration (slice-based) with normal scenario and other configuration (BWP-based) where the RAN was divided into two BWPs configured with the appropriate numerologies 0

and 1, with enabled mini-slot based scheduling and idle resource sharing capability, achieved KPI targets for all of the applications except the BB(UL) and VR application, respectively with closer target achieved with VR application with BWP-based configuration. With this observation, it is concluded and recommended that the same architecture choice and features as the best-performing configuration for normal scenario should be considered for all the cells for incident scenario, with a reconfiguration in incident cell to other BWP-based configuration when an incident occurs.

5.2. Recommendations for future work

Based on the investigation and overall analysis of the different 5G RAN features and their combinations in managing and supporting the three service categories in a smart city environment, some recommendations for the future work are suggested and are listed below.

- This study only considers one packet scheduler for all the configurations, which considers both latency and throughput aspects. It is suggested to use different packet schedulers specifically with the slicing configurations, which are dedicated to improve the performance of the slice-specific requirements. Also, it is recommended to consider a more advanced, suitably differentiating scheduler to challenge to the *need* for slicing as a differentiating mechanism.
- Considering the influence of the resource split ratio between slices/BWPs and the DL/UL ratio for determining the TDD configuration in the above results, a study should be carried out with an optimized resource split ratio between BWPs/slices and optimal TDD configurations. Also, it is recommended to consider the regulatory restrictions on the TDD split, a common TDD split for all the operators.
- In this study, static traffic is considered, with variability on a very fine scale, which is handled by the scheduling mechanism and/or by idle resource sharing capability. A study should be carried out with higher variability in traffic, and the concept of dynamic resource assignment to BWPs/slices should be investigated, possibly applying AI/ML based slice management solutions.
- It is suggested to investigate the impact of these 5G RAN features in achieving the performance requirements of different kind of scenarios from different verticals, for example, healthcare, industry 4.0, agriculture, logistics and many more.

A Appendix

A.1. MI-ESM curves

The MI-ESM curves are used to find the effective SINR for the sub-band by mapping the derived averaged MI values over all the PRBs in the subband, as mentioned in Section 3.6.1. The MI-ESM curves are shown in Figure A.1.

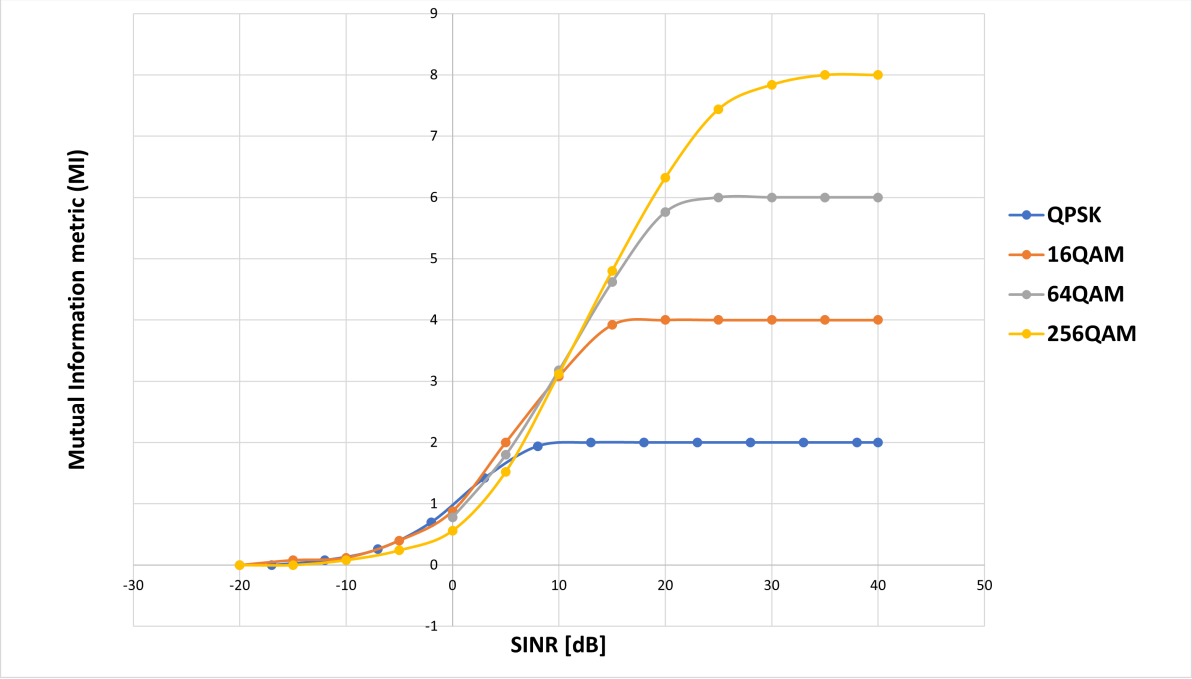

Figure A.1: The MI-ESM curves for mapping MI values to find effective SINR for the subband.

A.2. BLER curves

The BLER curves are used to find the maximum SINR value for each CQI value to achieve the target BLER based on the device type. For URLLC device, the target BLER is set to not exceed 0.001% and for eMBB and mMTC device, the target is set to not exceed 0.1%. The BLER curves used for URLLC devices are shown in Figure A.2 and for eMBB and mMTC device is shown in Figure A.3. These BLER curves are derived using the Vienna 5G Link Level Simulator [44].

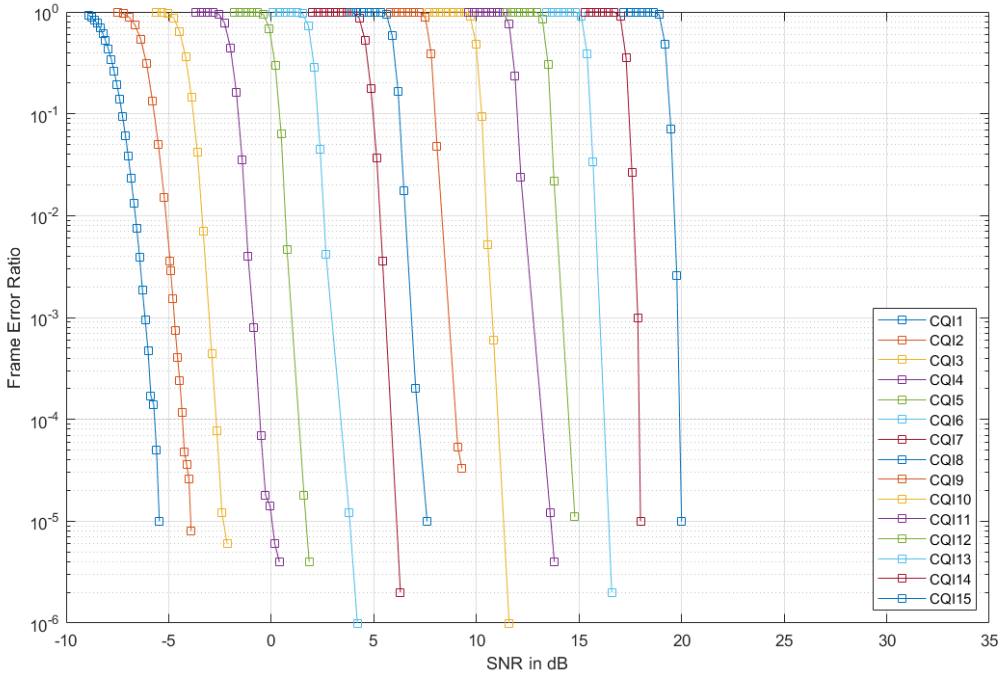

Figure A.2: The BLER curves for AWGN channel for each CQI value 0-15, for URLLC devices.

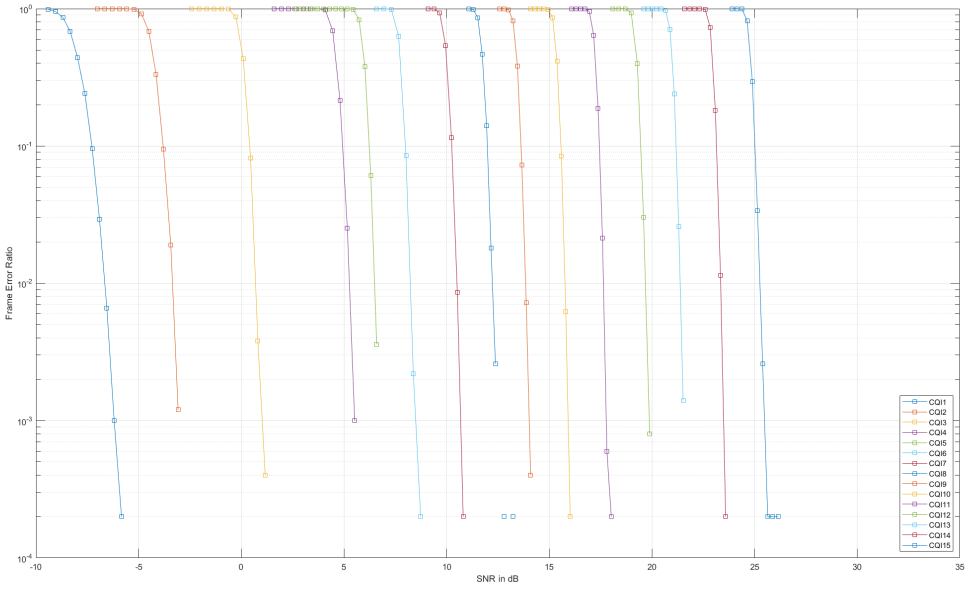

Figure A.3: The BLER curves for AWGN channel for each CQI value 0-15, for eMBB and mMTC devices.

Bibliography

[1] J. Navarro-Ortiz, P. Romero-Diaz, S. Sendra, P. Ameigeiras, J. J. Ramos-Munoz and J. M. Lopez-Soler, "A survey on 5G usage scenarios and traffic models," *IEEE Communications Surveys & Tutorials*, 2020.

[2] S. K. Rao and R. Prasad, "Impact of 5G technologies on smart city implementation," *Wireless Personal Communications*, 2018.

[3] D. Jiang and G. Liu, "An overview of 5G requirements," *5G Mobile Communications*, 2017.

[4] T. M. Ho, T. D. Tran, T. T. Nguyen, S. Kazmi, L. B. Le, C. S. Hong and L. Hanzo, "Next-generation wireless solutions for the smart factory, smart vehicles, the smart grid and smart cities," *arXiv preprint arXiv:1907.10102*, 2019.

[5] 5G-Americas, "5G communication for automation in vertical domains," 2018.

[6] R. El Hattachi and J. Erfanian, "NGMN 5G Initiative," *NGMN 5G White Paper v1.0*, 2015. [Online]. Available: `https://www.ngmn.org/wp-content/uploads/NGMN_5G_White_Paper_V1_0.pdf`.

[7] J. Morais, S. Braam, R. Litjens, S. Kizhakkekundil and J. van den Berg, "Performance modelling and assessment for social VR conference applications in 5G radio networks," *17th International Conference on Wireless and Mobile Computing, Networking and Communications*, 2021.

[8] NGMN-Alliance, "Perspectives on vertical industries and implications for 5G," *NGMN 5G White Paper*, 2016.

[9] R. Razavi, M. Fleury and M. Ghanbari, "Low-delay video control in a personal area network for augmented reality," *IET Image Processing*, 2008.

[10] Nokia, "An overview of network slicing for 5G," *IEEE Wireless Communications*, 2019.

[11] S. A. Kazmi, L. U. Khan, N. H. Tran and C. S. Hong, "Network slicing: The concept," *Network Slicing for 5G and Beyond Networks*, 2019.

[12] Nokia, "Transport slice controller application help," *Network services Platform*, 2021.

[13] A. A. Zaidi, R. Baldemair, V. Molés-Cases, N. He, K. Werner and A. Cedergren, "OFDM numerology design for 5G new radio to support IoT, eMBB, and MBSFN," *IEEE Communications Standards Magazine*, 2018.

[14] 3GPP, "NR and NG-RAN overall description," *Technical Specification Group Radio Access Network, TS 38.300 v15.8.0*, 2019.

[15] M. M. Teixeira, M. J. Santana and R. H. Santana, "Using adaptive priority scheduling for service differentiation QoS-aware web servers," *IEEE International Conference on Performance, Computing, and Communications*, 2004.

[16] S. E. Elayoubi, S. B. Jemaa, Z. Altman and A. Galindo-Serrano, "5G RAN slicing for verticals: Enablers and challenges," *IEEE Communications Magazine*, 2019.

[17] P. Rost, C. Mannweiler, D. S. Michalopoulos, C. Sartori, V. Sciancalepore, N. Sastry, O. Holland, S. Tayade, B. Han, D. Bega, *et al.*, "Network slicing to enable scalability and flexibility in 5G mobile networks," *IEEE Communications magazine*, 2017.

[18] C. Sexton, N. Marchetti and L. A. DaSilva, "Customization and trade-offs in 5G RAN slicing," *IEEE Communications Magazine*, 2019.

[19] J. Li, W. Shi, P. Yang, Q. Ye, X. S. Shen, X. Li and J. Rao, "A hierarchical soft RAN slicing framework for differentiated service provisioning," *IEEE Wireless Communications*, 2020.

[20] 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," *2018 IEEE 87th Vehicular Technology Conference (VTC Spring)*, 2018.

[21] S. Khatibi and A. Jano, "Elastic slice-aware radio resource management with AI-traffic prediction," *2019 European Conference on Networks and Communications (EuCNC)*, 2019.

[22] K. I. Pedersen, G. Pocovi, J. Steiner and S. R. Khosravirad, "Punctured scheduling for critical low latency data on a shared channel with mobile broadband," *2017 IEEE 86th Vehicular Technology Conference (VTC-Fall)*, 2017.

[23] 3GPP, "5G; System architecture for the 5G system," *TS 23.501 v16.6.0*, 2020.

[24] M. Raftopoulou and R. Litjens, "Optimisation of numerology and packet scheduling in 5G networks: To slice or not to slice?" *2021 IEEE 93rd Vehicular Technology Conference (VTC2021-Spring)*, 2021.

[25] 3GPP, "NR;physical channels and modulation," *TS 38.211 v16.3.0*, 2020.

[26] Erricson, "Mixed numerology in an OFDM system," *3GPP TSG RAN WG1 Meeting*, 2016.

[27] 3GPP, "User Equipment (UE) radio transmission and reception," *Technical Specification Group Radio Access Network, TS 38.101 v15.3.0*, 2019.

[28] 3GPP, "5G; NR; User Equipment(UE) radio transmission and reception," *TS 38.101-1 v15.3.0*, 2018.

[29] 3GPP, "5G; NR; Requirements for support of radio resource management," *TS 38.133 v15.3.0*, 2018.

[30] 3GPP, "Study on scenarios and requirements for next generation access technologies," *TS 38.913 v15.3.0*, 2018.

[31] GSMA, "5G TDD synchronisation," 2020. [Online]. Available: `https://www.gsma.com/spectrum/wp-content/uploads/2020/04/3.5-GHz-5G-TDD-Synchronisation.pdf`.

[32] R. Williamson, G. D'Aria and R. YN Li, "5G TDD Uplink," *NGMN 5G White Paper v1.0*, 2021. [Online]. Available: `https://www.ngmn.org/wp-content/uploads/220117-5G-TDD-Uplink-White-Paper-v1.0.pdf`.

[33] Erricson, "UP latency in NR," *3GPP TSG-RAN WG2 Meeting R2-1711550*, 2017.

[34] G. C. Buttazzo, M. Bertogna and G. Yao, "Limited preemptive scheduling for real-time systems. a survey," *IEEE transactions on Industrial Informatics*, 2012.

[35] 3GPP, "Study on channel model for frequencies from 0.5 to 100 ghz," *Technical Report TR 38.901 v14.0.0*, 2019.

[36] F. Gunnarsson, M. N. Johansson, A. Furuskar, M. Lundevall, A. Simonsson, C. Tidestav and M. Blomgren, "Downtilted base station antennas- a simulation model proposal and impact on HSPA and LTE performance," *2008 IEEE 68th Vehicular Technology Conference*, 2008.

[37] S. Jaeckel, Raschkowski, L. Thiele, F. Burkhardt and E. Eberlein, "QuaDRiGa - Quasi deterministic radio channel generator: User manual and documentation," *Fraunhofer Heinrich Hertz Institute, Technical Report V2. 4.0*, 2020.

[38] P. Sarigiannidis, M. Louta and A. Michalas, "On effectively determining the downlink-to-uplink sub-frame width ratio for mobile wimax networks using spline extrapolation," *2011 15th Panhellenic Conference on Informatics*, 2011.

[39] S.-C. Tseng, Z.-W. Liu, Y.-C. Chou and C.-W. Huang, "Radio resource scheduling for 5G NR via deep deterministic policy gradient," *2019 IEEE International Conference on Communications Workshops (ICC Workshops)*, 2019.

[40] 3GPP, "Traffic models for xr," *3GPP TSG RAN WG1 e-meeting*, 2021. [Online]. Available: `https://www.3gpp.org/ftp/TSG_RAN/WG1_RL1/TSGR1_104-e/Docs/R1-2101493.zip`.

[41] S. research department, "Average residential property floor space in selected european countries in 2014," *Statista*, 2014. [Online]. Available: `https://www.statista.com/statistics/506043/average-floor-space-homes-in-europe/`.

[42] 3GPP, "5G NR; physical layer procedures for data," *TS 38.214 v15.3.0*, 2018.

[43] K. Sayana, J. Zhuang and K. Stewart, "Link performance abstraction based on mean mutual information per bit (mmib) of the llr channel," *IEEE 802.16 Broadband Wireless Access Working Group*, 2007.

[44] S. Pratschner, B. Tahir, L. Marijanovic, M. Mussbah, K. Kirev, R. Nissel, S. Schwarz and M. Rupp, "Versatile mobile communications simulation: The vienna 5G link level simulator," *EURASIP Journal on Wireless Communications and Networking*, vol. 2018, 2018.

[45] S. N. Anbalagan, R. Litjens, K. Das, A. Chiumento, P. Havinga and H. van den Berg, "A sensitivity analysis on the potential of 5G channel quality prediction," *2021 IEEE 93rd Vehicular Technology Conference (VTC2021-Spring)*, 2021.

[46] F. Capozzi, G. Piro, L. A. Grieco, G. Boggia and P. Camarda, "Downlink packet scheduling in lte cellular networks: Key design issues and a survey," *IEEE communications surveys & tutorials*, 2012.

[47] P. Baracca, L. G. Giordano, A. Garcia-Rodriguez, G. Geraci and D. López-Pérez, "Downlink performance of uplink fractional power control in 5G massive mimo systems," *2018 IEEE Globecom Workshops (GC Workshops)*, 2018.

[48] M. Raftopoulou, "Design and assessment of random access procedures supporting massive connectivity and low-delay and high-reliability services in 5G," 2018.

[49] 3GPP, "5G; NR; User Equipment(UE) radio transmission and reception," *TS 38.101-1 v16.4.0*, 2020.