Dept. of Communications Engineering

5G NR Physical Layer Design for Ultra-Reliable Low-Latency Communication

Master Thesis

submitted by

Venkateswarlu Yampati

Bremen, January 13, 2023

University of Bremen

5G NR Physical Layer Design for Ultra-Reliable Low-Latency Communication

Fachbereich 1 - Physics and Electronical Engineering
Institute for Telecommunications and High-Frequency Techniques (ITH)
Department of Communications Engineering
P.O. Box 33 04 40
D-28334 Bremen

Supervisor:	Fayad Haddad, M.Sc.
First Examiner:	Prof. Dr.-Ing. A. Dekorsy
Second Examiner:	Dr.-Ing. C. Bockelmann

I ensure the fact that this thesis has been independently written and no other sources or aids, other than mentioned, have been used.

Bremen, January 13, 2023 Y. Venkateswarlu

.....

Universität Bremen

Hinweise zu den offiziellen Erklärungen

1. Alle drei Erklärungen sind **unverändert im Wortlaut** in **jedes Exemplar** der BA-/MA-Arbeit **fest mit einzubinden** und jeweils **im Original zu unterschreiben**.

2. In der digitalen Fassung kann auf die Unterschrift verzichtet werden. Die Angaben und Entscheidungen müssen jedoch enthalten sein.

Zu A

Bitte ergänzen Sie die notwendigen Angaben.

Zu B

Die Einwilligung kann jederzeit durch Erklärung gegenüber der Universität Bremen mit Wirkung für die Zukunft widerrufen werden.

Zu C

Das Einverständnis mit der Überprüfung durch die Plagiatsoftware *Plagscan* und der dauerhaften Speicherung des Textes ist freiwillig. Die Einwilligung kann jederzeit durch Erklärung gegenüber der Universität Bremen mit Wirkung für die Zukunft widerrufen werden.

Im Jahr 2019 wird die Software zunächst in einigen Fachbereichen eingesetzt.

Weitere Informationen zur Überprüfung von schriftlichen Arbeiten durch die Plagiatsoftware sind im Nutzungs- und Datenschutzkonzept enthalten. Diese finden Sie auf der Internetseite der Universität Bremen.

Universität Bremen

Offizielle Erklärungen von

Name: Venkateswarlu Yampati Matrikelnr.: 3201063

Eigenständigkeitserklärung
Ich versichere, dass ich die vorliegende Arbeit selbstständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel verwendet habe.
Alle Teile meiner Arbeit, die wortwörtlich oder dem Sinn nach anderen Werken entnommen sind, wurden unter Angabe der Quelle kenntlich gemacht. Gleiches gilt auch für Zeichnungen, Skizzen, bildliche Darstellungen sowie für Quellen aus dem Internet.
Die Arbeit wurde in gleicher oder ähnlicher Form noch nicht als Prüfungsleistung eingereicht.
Die elektronische Fassung der Arbeit stimmt mit der gedruckten Version überein.
Mir ist bewusst, dass wahrheitswidrige Angaben als Täuschung behandelt werden.

A) Erklärung zur Veröffentlichung von Bachelor- oder Masterarbeiten

Die Abschlussarbeit wird zwei Jahre nach Studienabschluss dem Archiv der Universität Bremen zur dauerhaften Archivierung angeboten. Archiviert werden:
1) Masterarbeiten mit lokalem oder regionalem Bezug sowie pro Studienfach und Studienjahr 10 % aller Abschlussarbeiten
2) Bachelorarbeiten des jeweils ersten und letzten Bachelorabschlusses pro Studienfach u. Jahr.

☒ Ich bin damit einverstanden, dass meine Abschlussarbeit im Universitätsarchiv für wissenschaftliche Zwecke von Dritten eingesehen werden darf.

☒ Ich bin damit einverstanden, dass meine Abschlussarbeit nach 30 Jahren (gem. §7 Abs. 2 BremArchivG) im Universitätsarchiv für wissenschaftliche Zwecke von Dritten eingesehen werden darf.

☐ Ich bin nicht damit einverstanden, dass meine Abschlussarbeit im Universitätsarchiv für wissenschaftliche Zwecke von Dritten eingesehen werden darf.

B) Einverständniserklärung über die Bereitstellung und Nutzung der Bachelorarbeit / Masterarbeit / Hausarbeit in elektronischer Form zur Überprüfung durch Plagiatssoftware

Eingereichte Arbeiten können mit der Software *Plagscan* auf einem hauseigenen Server auf Übereinstimmung mit externen Quellen und der institutionseigenen Datenbank untersucht werden.
Zum Zweck des Abgleichs mit zukünftig zu überprüfenden Studien- und Prüfungsarbeiten kann die Arbeit dauerhaft in der institutionseigenen Datenbank der Universität Bremen gespeichert werden.

☒ Ich bin damit einverstanden, dass die von mir vorgelegte und verfasste Arbeit zum Zweck der Überprüfung auf Plagiate auf den *Plagscan*-Server der Universität Bremen hochgeladen wird.

☒ Ich bin ebenfalls damit einverstanden, dass die von mir vorgelegte und verfasste Arbeit zum o.g. Zweck auf dem *Plagscan*-Server der Universität Bremen hochgeladen u. dauerhaft auf dem *Plagscan*-Server gespeichert wird.

☐ Ich bin nicht damit einverstanden, dass die von mir vorgelegte u. verfasste Arbeit zum o.g. Zweck auf dem *Plagscan*-Server der Universität Bremen hochgeladen u. dauerhaft gespeichert wird.

Mit meiner Unterschrift versichere ich, dass ich die obenstehenden Erklärungen gelesen und verstanden habe. Mit meiner Unterschrift bestätige ich die Richtigkeit der oben gemachten Angaben.

13.01.2023, Bremen — Y. Venkateswarlu

Datum, Ort — Unterschrift

Contents

Chapter 1

Introduction

1.1 Motivation

A new class of use cases and applications, such as high throughput, massive device connections, augmented reality (AR), virtual reality (VR), industrial automation, autonomous vehicles, etc., have prompted the third-generation partnership project (3GPP) to specify different service categories. The 3GPP specified three service paradigms for fifth-generation (5G) new radio (NR) cellular networks: enhanced mobile broadband (eMBB), massive machine-type communication (mMTC), and ultra-reliable low-latency communications (URLLC). eMBB sets a goal for 5G to achieve higher peak download rates, which is far faster than the peak download speed that the fourth generation (4G) long term evolution (LTE) can deliver. mMTC is used to establish the minimum criteria for 5G to support the huge number of low-powered, low-cost, low-complexity devices per square kilometer with a longer battery life. The URLLC service, by definition, requires high reliability and low latency communications. There are several benefits to autonomous driving, from time savings to greater safety by decreasing human errors. For instance, autonomous driving would need a connection that can offer this service given the high level of danger involved. URLLC provides use cases that demand high network reliability of more than 99.999 percent and exceptionally low latency of just 1 ms for data transmission [1]. All vehicles would need to be linked up vehicle-to-vehicle (V2V) and vehicle-to-infrastructure (V2X), such as for emergency services, traffic signal systems, and road maintenance initiatives. Due to the need for ultra-reliable connections, data would need to be transferred in real time with little delay. Similar criteria apply to Industry 4.0 and smart factories, where machinery and robotics must communicate with one another in real-time. Additionally, they could need real-time data from various sensors located throughout the manufacturing site. These machine-operated systems can run securely and effectively to improve production lines due to low-latency solutions. As a result, cellular wireless communications must meet these requirements for high reliability and low latency.

The physical layer (PHY) in the present 4G LTE wireless network is not viable for this kind of communication because it is not designed specifically for achieving high reliability and latency requirements. Hence, in Release 15, the first complete set of 5G physical layer designs for URLLC, was introduced by 3GPP [2]. 3GPP is a partnership project that standardizes cellular telecommunications technologies that give a comprehensive system description for mobile telecommunications, such as radio access, the core network, and service capabilities. Technical specifications or releases

and reports are produced by 3GPP for both the Time Division Duplex (TDD) and Frequency Division Duplex (FDD) radio access technologies. Release 16, the second 5G release, was completed in December 2019 and enables increased latency and reliability measurements to accommodate new URLLC use cases. To address new industrial use cases, 3GPP Release 17 expands URLLC functionalities to unlicensed spectrum. This thesis focuses on physical layer design for 5G URLLC. The resource management and allocation in both uplink and downlink for a network configured with URLLC users and the presence of eMBB users are presented based on physical layer enhancements proposed in releases 15 and 16.

1.2 Contribution of the Thesis

This thesis discusses the physical layer design for applications demanding ultra-reliability and low latency. Also, it discusses the physical layer resource allocation with the multiplexing of two services: eMBB and URLLC. The contribution is as follows:

- Discussion of physical layer enhancements to enhance reliability and latency in both uplink and downlink from 3GPP releases 15 & 16.
- Implementation of a resource allocation optimization algorithm for URLLC users in the downlink.
- Implementation of downlink resource allocation for both URLLC and eMBB users in the network.
- A comparison between various uplink resource allocation schemes for URLLC service with 3GPP releases.
- Implementation of an uplink resource allocation algorithm for both URLLC and eMBB users to achieve their service requirements respectively.
- A comparison of implemented optimized scheduling algorithms with baseline scheduling algorithms in the literature.

1.3 Thesis Organization

Chapter 1: This chapter motivates the requirement of highly reliable and low-latency communications for 5G NR.

Chapter 2: This chapter provides the summary of 5G NR, which includes the description of three service categories introduced by 3GPP in Release 15. The definitions of latency and reliability and the factors affecting them are also presented.

Chapter 3: This chapter gives the concept of the physical layer in 5G NR and the enhancements or features introduced in 3GPP Releases 15 and 16 to improve reliability and decrease latency for the applications of URLLC. The definitions of baseline scheduling algorithms introduced in wireless cellular communications are also presented in this chapter.

Chapter 4: This chapter discusses the downlink resource allocation of URLLC users as well as the problem of resource allocation when both URLLC and eMBB users coexist in the network. By converting the resource allocation problem to a convex optimization problem with latency constraints, the resources in the network are assigned to different services or users. The formulated resource allocation problem is subject to reliability and latency constraints for URLLC users. To ensure the reliability of eMBB and URLLC services, a modulation scheme which is adaptive in nature also known as adaptive modulation scheme (AMC) is used. The resource allocation for both URLLC and eMBB users is presented based on the data sum-rate maximization of all users in the downlink, with reliability constraints for both types of services and latency constraints for the URLLC service. Finally, a comparison between the proposed scheduling algorithm and baseline algorithms in the literature is presented.

Chapter 5: The time for initial access to the resources from the NR base station also called as next generation NodeB (gNB) plays a vital role in latency for URLLC services. One of the physical layer enhancements introduced in 3GPP Release 15 is the ability to allocate resources to users in the uplink without requiring a grant from the base station, reducing initial access latency. Even though, this enhancement decreases the latency, it has an impact on reliability. 3GPP proposed diversity based transmission i.e, repetitions of same data in certain time interval to ensure the reliable data transmissions. In this chapter, the concept of contention-based access, which is introduced for grant-free (GF) transmissions to decrease the latency in the uplink, is discussed. Following it, the optimal resource allocation scheme in the uplink are compared with 3GPP Release 15 and 16 solutions. Finally, resource allocation in uplinks for both eMBB and URLLC users is discussed using the same optimization techniques discussed in above chapter. To the best of our knowledge, this is the first time that the resource allocation in the uplink for eMBB and URLLC services is presented by considering the optimal resource allocation scheme for URLLC users (ensuring repetitions of packet following Rel 16) and thereby maximizing the data sum-rate for eMBB users with the AMC scheme.

1.3.1 Notations and Nomenclature

Following is a brief list of the nomenclature that is used.

- The numbers and scalars are denoted as italic uppercase letters (e.g. F) and italic lower case letters (e.g. f) respectively.
- The matrices are represented with bold upper case ($\mathbf{F}$) and vectors are represented with bold lower case letters($\mathbf{f}$).
- Log represents the natural logarithm (base e), whereas log2 represents the logarithm to base 2.
- $||k||^2$ always refers to the l_2-norm of a matrix, whereas $|k|$ specifies the absolute value of a variable.

,

Chapter 2

5G New Radio System

On top of traditional cellular mobile broadband services, the 5G NR wireless communication technology is projected to accommodate a wide range of new upcoming applications. The 5G NR ultra-high-speed wireless communication standard represents a significant technological advance, significantly improving speed and capacity, expanding current use cases, and enabling many new applications. The current 4G LTE systems were designed to mainly support human-centric applications such as voice calls and mobile broadband connectivity. However, emerging applications like VR / AR, Internet of Things (IoT), V2X, etc., require higher data rates, massive connectivity, and fast and high accuracy [3]. In Release 15, the 3GPP has defined three 5G service categories to meet communication requirements for supporting planned applications. These are enhanced mobile broadband (eMBB), massive machine type communications (mMTC), and ultra-reliable and low latency (URLLC). These services support several applications, as shown in Fig. 2.1.

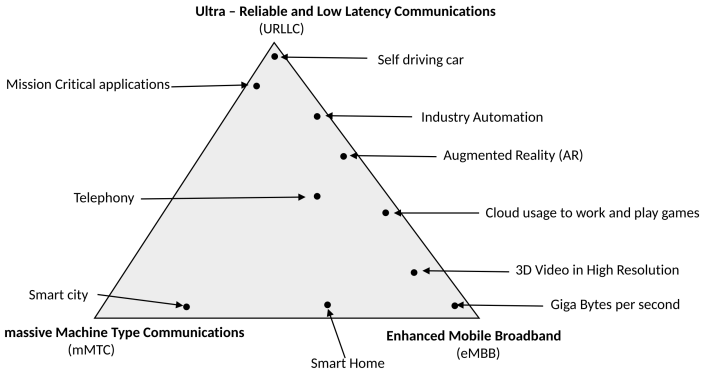

Figure 2.1: 5G services and their applications

2.1 enhanced Mobile Broad Band (eMBB)

The human-centric use cases for accessing multimedia content, services, and data are addressed by eMBB. eMBB is a data transmission service that meets the demand for high data rates. Demand for mobile broadband will keep rising, resulting in improved mobile broadband. Present 4G LTE networks will naturally evolve to eMBB, which will offer higher data speeds and, consequently, a better user experience than current mobile broadband services[1]. It is an advanced type of 4G that offers better performance and a more smooth user experience that goes beyond just quicker downloads. In the end, it will make it possible for fully immersive VR, AR applications, and online gaming, which require higher bandwidth [4]. Indeed, the demand for higher data rates drove the development of previous generations of cellular systems, including 3G and 4G.

The key capabilities that the eMBB service focuses on are data throughputs, spectral efficiency, mobility, and area traffic capacity. Several physical layer technologies, like high-order modulation transmission, carrier aggregation (CA), and multiple input multiple output (MIMO) transmissions are introduced in 4G standards for higher user experience [1]. However, the increasing traffic demand and competition among operators drove the requirement for higher data rates and high quality of service (QoS). It is believed that in unfavorable situations, such as highly populated places with high mobility, 5G promises to provide a peak data throughput of 20 gigabits per second (Gbps) while still ensuring a good data rate (50–100 Mbps) and ubiquity [1]. In addition, the target delay for user data transfer should also be minimal. Physical layer technologies that can enhance the throughput under consideration are massive MIMO [5] and millimeter-wave communication (mmWave) [6].

2.2 massive Machine Type Communications (mMTC)

mMTC is introduced in 5G NR to support a large number of device communications. It was developed primarily to make it possible to simultaneously gather a significant amount of tiny data packets from numerous devices. This service can provide solutions to IoT by providing efficient connectivity with the network and among devices. Asset tracking, smart agriculture, smart cities, energy monitoring, smart homes, and remote monitoring are a few examples of use cases in this category. This service can handle connection densities of up to one million devices per square kilometer [7]. This is ten times more capable than a 4G LTE network. This capability enables 5G NR to deliver the foundation required to support massive networks of cellular-connected sensors. These IoT gadgets must have a lengthy battery life, possibly lasting up to 10 years [7].

mMTC-based services, such as sensing, tagging, metering, and monitoring, require high connection density and better energy efficiency [7]. On top of the huge connectivity, these applications may require additional communication requirements. Some of the devices, for example, are battery-powered and should last for several years without needing to be recharged batteries. Other devices, such as smart meters, may be used in areas where penetration loss is considerable. As a result, they require greater network coverage for their operations. Furthermore, for many IoT applications, a low-cost transceiver is required.

2.3 Ultra-Reliable Low-Latency Communication (URLLC)

URLLC is a service category to support latency-sensitive services in which the packet transmission time must be very minimal. Remote control, automatic driving, and tactile internet are examples of these services. These services have stringent requirements on availability, latency, and reliability [8]. Devices requiring low latency and good link reliability can benefit from URLLC. That implies that the communication link between the devices must be available all the time to carry the data in a short amount of time with high accuracy. Many automobile factories are transitioning to increase production efficiency due to the introduction of autonomous and collaborative driving. For these modifications to be possible, URLLC is a crucial component. There are countless URLLC applications, including as

- Health care: a surgeon in another location can do remote surgery with the aid of a robot that gets instructions in real time.
- V2X: a vehicular communication system that facilitates the transmission of data from a moving vehicle to other moving traffic-related elements that could have an impact on the moving vehicle.
- AR: refers to a technologically improved version of the real world that uses digital visual components, sound, or other sensory stimulation.
- Tactile Interaction: the next iteration of the IoT that combines exceptionally high availability, reliability, and security with extremely low latency.
- Self - driving car: is a category of vehicle that integrates vehicular automation, which is the ability of a ground vehicle to perceive its environment and maneuver safely with little or no human interaction.
- Smart grid: an electrical grid controlled remotely or automatically by automatic control mechanisms.
- Motion control: the isochronous transfer of sensory and actuation information in the uplink and downlink, which demands a latency $\leq$ 1 ms, and real-time control of machines with moving parts.
- Factory automation: is an integrated industrial process that automates operations, processes, and production using a variety of technologies in order to increase output and efficiency while lowering total costs.
- Process Automation: Automation often reduces the amount of thought or effort required from a person to complete a task with an end-to-end (E2E) latency of 50 ms.

The Table. 2.1 represents the use cases and their requirements for URLLC in 5G NR [4].

Table 2.1: URLLC use cases and requirements

Scenario	**E2E (ms)**	**Reliability (%)**
Discrete automation-motion control	1	99.9999
Electricity distribution-high voltage	5	99.9999
Remote Control	5	99.999
Discrete automation	10	99.99
Intelligent transport systems	10	99.9999
Process automation	50	99.9999
Electricity distribution	25	99.9

2.4 URLLC Overview

The design of the physical layer for URLLC is the most difficult of the three service types defined by 3GPP because it must simultaneously address the two competing considerations of reliability and latency. These two performance indicators are dependent on each other. To increase the reliability of the transmitted data packet, the retransmissions of the undecoded or corrupted data can be performed but this leads to an increase in latency. This section explains the definitions of the key performance indicators (KPI) for the URLLC applications and the factors affecting them.

2.4.1 Latency

Latency is the radio network's contribution to the delay between the time the source sends a packet and the time the destination receives it, measured in milliseconds (ms). It is determined by the one-way time required to successfully transmit an application layer packet or message from the radio protocol layer 2/3 service data unit (SDU) ingress point to the radio protocol layer 2/3 SDU egress point of the radio interface [9]. The SDU in a packet is information that the protocol sends between peer protocol entities on behalf of the users who use the services provided by that layer. In 5G, layer 1 is the physical layer (PHY), layer 2 is comprised of medium access control (MAC), radio link control (RLC), and packet data convergence protocol (PDCP), and layer 3 is the radio resource control layer (RRC) [10]. Many features are available in 3GPP for all layers. However, in this thesis, the discussion of reducing PHY layer latency is considered as shown in Fig. 2.2. The latency of the physical layer is comprised of the following five components, as shown in Fig. 2.3:

$$T_L = T_t + T_p + T_e + T_r + T_s, \tag{2.1}$$

- T_t is time to transmit a packet.
- T_p is the time for the signal to propagate from transmitter to the receiver.
- T_e is the time to perform encoding and decoding and also the channel estimation in the initial transmission.

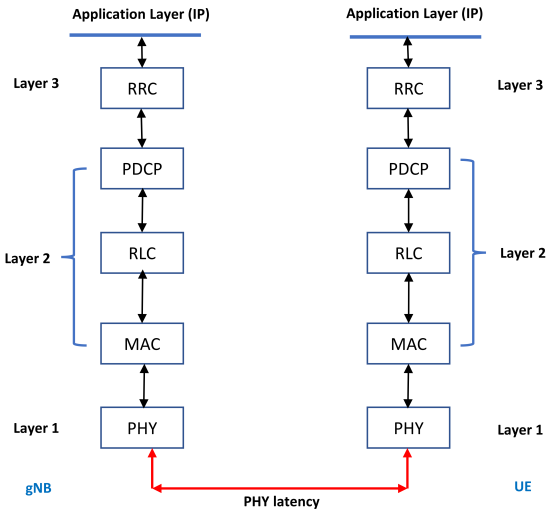

Figure 2.2: 5G protocol stack

- T_r is the re-transmission time in case when the sender received Negative acknowledgment (NACK) or acknowledgement (ACK) from receiver.
- T_s is the pre-processing time for the signal exchange such as connection request, scheduling grant, channel training and feedback and queuing delay.

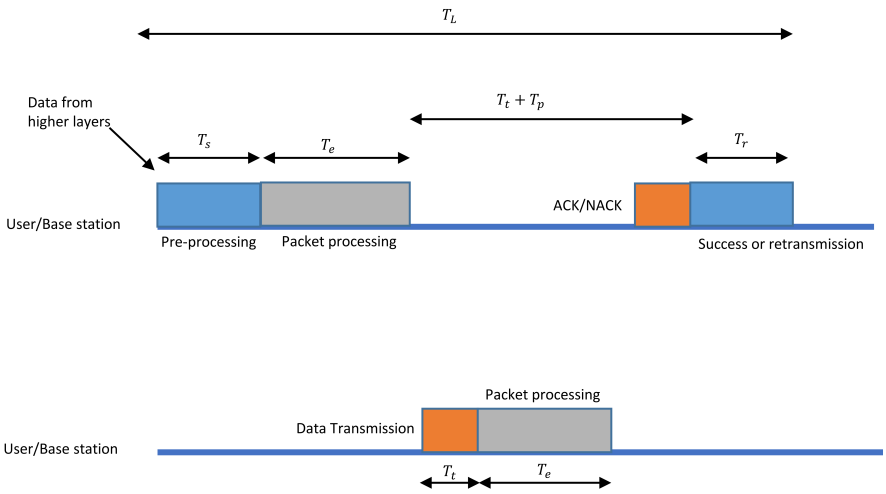

Figure 2.3: Latency components

The time to transmit a packet is determined by the sub-carrier spacing (SCS) used. In LTE

systems, it is fixed to 15 kHz, hence the time to transmit a packet (T_t) is 1 ms [1]. But the URLLC requirement is very stringent, such that a packet transmission time T_t should be in the hundreds of microseconds (μs) range to achieve the high latency limitation. Hence, in 5G NR, the usage of various sub-carrier spacings is introduced, thereby decreasing the latency. The relation between SCS and packet transmit time is presented in the next chapter. The latency is also impacted by the initial connection setup (T_s) between sender and receiver and the re-transmission time (T_r) in case of errors. In 5G NR Release 15, the technique of uplink grant-free transmissions is introduced to reduce the initial connection setup delay, thereby reducing the latency [11]. A brief description of these enhancements based on 3GPP standards is discussed in the next chapter.

2.4.2 Reliability

Reliability is considered an important performance indicator for the URLLC service. When a small data packet needs to be delivered from the radio protocol layer 2/3 SDU ingress point to the radio protocol Layer 2/3 SDU egress point of the radio interface at a specific channel quality, the probability that it will do so successfully and within the required maximum time is known as reliability [4]. The 3GPP defines URLLC requirements to support use cases such as smart grid, AR, and VR in entertainment industry: "A general URLLC reliability requirement for one transmission of a packet is 10^{-5} for 32 bytes with a user-plane latency of 1 ms" [1]. This criterion presents a challenge in URLLC design because it is significantly higher than the normal block error rate (BLER) of a LTE system, which is 10^{-1}. URLLC additions in Release 16 have revised the requirements even more, setting a reliability objective of 10^{-6} and a latency range of 0.5 to 1 ms to accommodate new use cases such as factory automation, the transportation industry, including remote driving, and electrical power distribution [12]. 3GPP introduced the AMC schemes to ensure the reliability of the communication link by using modulation techniques based on user equipment (UE) reported signal to noise (SNR) power to the base station in both uplink and downlink data transmissions. Like mentioned before, the UE can also perform repetitions of same data packet for the reliable data transmissions. These techniques introduced by 3GPP to increase reliability are discussed in the next chapter.

Chapter 3

Physical Layer Solutions for URLLC

This chapter will give an overview of the physical layer fundamentals, including orthogonal frequency division multiplexing (OFDM), time-frequency transmission resources in NR, resource scheduling, physical channels that carry information, and details about the enhancements in the physical layer for 5G URLLC. The physical layer serves as the foundation for 5G NR. From 1 Giga hertz (GHz) to 100 GHz, the NR physical layer must handle a wide variety of frequencies and different deployment techniques. The 3GPP is creating a flexible physical layer for NR in order to effectively handle different difficulties faced in future mobile communications.

3.1 Orthogonal Frequency Division Multiplexing

OFDM is a digital data modulation technique used in data communications and networking that splits a single data stream into numerous distinct sub-streams for transmission over a number of channels. The frequency division multiplexing (FDM) principle, which divides the available bandwidth into a number of sub-streams with distinct frequency bands, is the basis for OFDM. The common problem in wireless communication is frequency-selective fading, in which various frequency components are faded differently by the channel. In order to reduce this effect, OFDM was introduced [13]. In addition to requiring orthogonality between sub-channel signals, OFDM is a unique instance of multi-carrier modulation. Consequently, OFDM can be viewed as a modulation system or a multiplexing approach. The multi-carrier transmitter typically comprises of a number of modulators, each with a unique carrier frequency (Appendix A). OFDM with multiple access (OFDMA) can support multiple users in the network by assigning different carriers. Hence the OFDMA is adapted in 4G LTE and 5G NR systems for data transmissions [13]. The carrier frequencies with same spacing are known as sub-carriers and space is known as sub-carrier spacing which was mentioned earlier. In LTE, the SCS is 15 kHz, and the SCS in 5G NR is scalable and described as $15 * 2^{\eta}$ kHz, where η is an integer. The η takes values 0,1,2,3 [14]. Based on integer η, the sub-carrier spacing changes. This variation of having different sub-carrier spacings based on chosen bandwidth is called numerology. The maximum bandwidths enabled by various numerologies in 5G are provided in Table 3.1 [14].

Table 3.1: Scalable OFDM numerology for 5G NR

OFDM numerology	**15 kHz**	**30 kHz**	**60 kHz**	**120 kHz**
OFDM symbol duration (μs)	66.67	33.33	16.67	8.33
Cyclic prefix (CP) duration (μs)	4.69	2.34	1.17	0.59
OFDM symbol with CP (μs)	71.35	35.68	17.84	8.91
Maximum bandwidth(MHz)	50	100	200	400

3.2 Physical Time-Frequency Resources

OFDM symbols and subcarriers within those symbols relate to actual time-frequency resources. NR transmissions are organized in the time domain into frames of length 10 ms, each of which is divided into 10 equally sized sub-frames of length 1 ms, and a sub-frame is further divided into slots consisting of 14 OFDM symbols each; thus, the length of a slot measured in ms is determined by the numerology as shown in Fig 3.1. An NR slot thus has the same structure as an LTE subframe for the 15 kHz subcarrier spacing, which is advantageous from a coexistence standpoint for deployments. The slot is the conventional dynamic scheduling unit, whereas a subframe in NR acts as a numerology-independent time reference. Given that the duration of an OFDM signal is inversely proportional to its subcarrier spacing, the time duration of a slot or mini-slot scales with the chosen numerology. In 5G NR, mini-slot scheduling of time-resources is also introduced which can be even one OFDM symbol long [15].

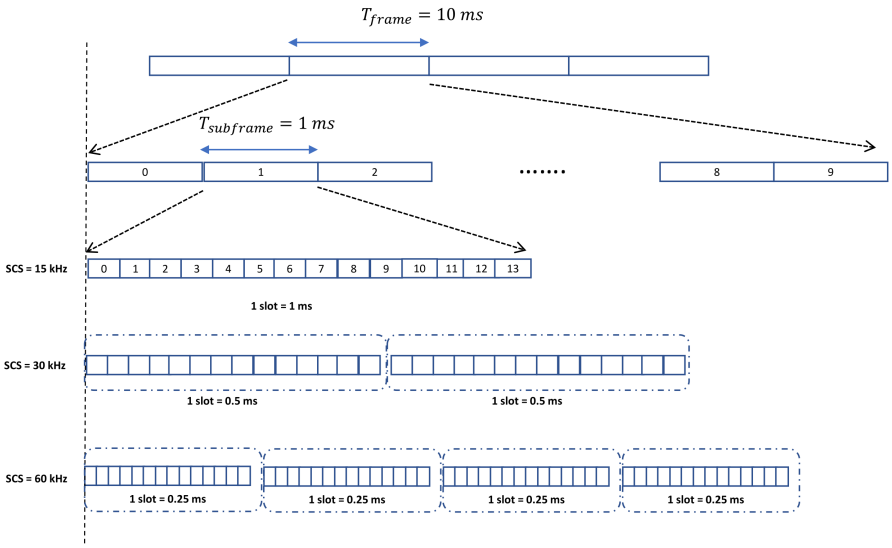

Figure 3.1: Slot structure with respect to sub-carrier spacing

The smallest physical time-frequency resource is one subcarrier in one OFDM symbol which is

defined as the resource element (RE). The physical resource block (PRB) or resource block (RB) in which the transmissions are scheduled are groups of 12 subcarriers. The NR physical time - resource frame structure is shown in Fig 3.2.

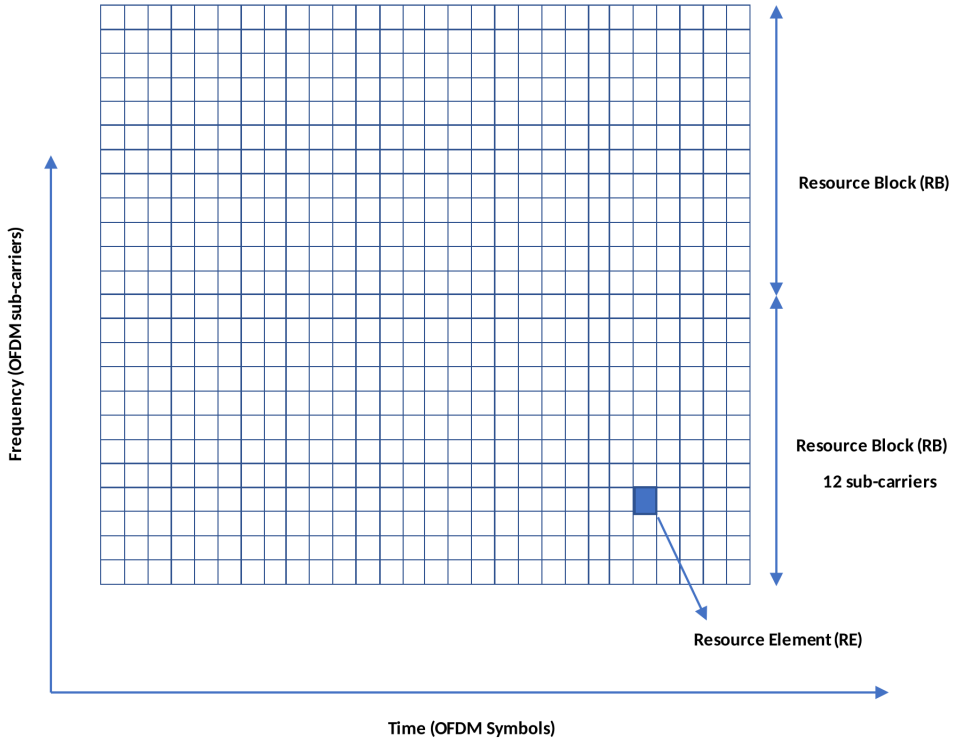

Figure 3.2: Physical time-frequency resources

3.3 Physical Layer Resource Allocation

The time-frequency resources discussed above are scheduled for users in the network. The base station uses various kinds of scheduling algorithms to allocate these resources both in time and frequency domain. The scheduling of these resources can be divided based on two main factors: channel aware and QoS aware scheduling algorithms [16]. Channel aware scheduling implies that the resource allocation is done based on channel conditions reported by UE. QoS aware resource allocation implies that the network has different service requirements demanding various KPI's like video, audio traffic. The 5G system is introduced to support the different service categories. The presence of multiple services in the network makes assigning resources subject to service requirements difficult. This section will explain the various resource allocation algorithms that are presented in literature.

3.3.1 Round Robin Resource Allocation

The easiest method of allocating resources is by using the round robin algorithm [17]. A list of active users is generated by the scheduler and is then sorted at random. RR is the simplest scheduling method used in wireless networks. A round-robin scheduler typically uses time-sharing, assigning each work a time slot, and interrupting the job if it is not finished by then, in order to schedule processes evenly. Each user receives an equitable share of the available resources. Scheduling using round-robin is straightforward, simple to implement, and starvation-free.

3.3.2 Proportional Fair Resource Allocation

A scheduling algorithm based on compromise is called proportional-fair (PF) scheduling [17]. It is centered on balancing two conflicting interests: attempting to optimize the total throughput of the network, while also allowing all users to receive at least a basic level of service. This means the proportional fairness scheduling algorithm is a packet scheduling technique that can achieve high overall throughput while still offering users a decent level of fairness[18] [16]. Cell throughput is an important measure to consider when comparing wireless packet schedulers due to the limited availability of wireless bandwidth. A scheduler that achieves the highest cell throughput is trivial to design. In this instance, if the channel gain reported by user e is higher than any other users in the network, then the base station served this user. This scheduler is considered as channel aware but not QoS aware scheduler. The PF scheduler strikes a fair balance by distributing an equal number of time slots to each user functioning as follows:

- Each user (e) provides the feasible rate (instantaneous) or channel status to the BS at the start of each time slot, represented as $r^e_{i,j}$, where i and j represent time and frequency domains, respectively.
- A moving average is used to monitor each user's (e) average throughput denoted by $T^e_{i,j}$. It is given in Eq. 3.1 [19]. Over a predetermined averaging period T_{PF}, the average is determined. The fixed size time window T_{PF} varies in such a way that the utility functions objective is either providing maximum throughput to all users or lower throughput while providing the certain throughput to all users in the network [16]. The time constant T_{PF} should be set to be longer than the time constant for the short-term fluctuations in order to enable effective use of the short-term channel variations. At the same time,T_{PF} should also be short enough so that a user won't notice quality changes significantly inside the T_{PF} interval.

$$T^e_{i,j} = (1 - \frac{1}{T_{PF}}) \cdot T^e_{i-1,j} + \frac{1}{T_{PF}} \cdot r^e_{i-1,j} \tag{3.1}$$

- The BS determines the possible rate to average throughput ratio for each user denoted as $\frac{r^e_{i,j}}{T^e_{i,j}}$. The user which gives the maximum value of this metric is scheduled by base station. This metric is known as the preference metric.
- The user with the highest preference metric will be chosen for transmission during the upcoming time period. Therefore, the objective or utility function to maximize the sum rate of E users is given by

$$U = \max \quad \sum_{e=1}^{E} \frac{[r_{i,j}^{e}]}{[T_{i,j}^{e}]} \tag{3.2}$$

3.3.3 Maximum Data Sum Rate Resource Allocation

This is also known as best SNR algorithm. On the basis of the best channel quality condition, the maximum scheduling algorithm schedules end users. Resources will be assigned to end users with good channel quality condition providing high data rate and cell edge users with low data rate. The utility function for the maximum data sum rate algorithms given as follows:

$$U = \max \quad \sum_{e=1}^{E} [r_{i,j}^{e}] \tag{3.3}$$

In this thesis, the 5G NR system must support two services: eMBB and URLLC. To check the performance comparisons, the total number of resources are divided equally between two different types of services (QoS based), then the scheduling algorithm is considered as equally distributed scheduler (EDS). The data rate comparisons of these base-line algorithms and the proposed algorithm is discussed in further sections.

3.4 Physical Channels

It is important to understand how the data in time-frequency resources is transmitted between sender and receiver. Physical channels are the time-frequency resources that transmit data from higher levels (layers above physical layer) [4]. For both uplink and downlink, the physical channels as specified as follows:

- Physical downlink control channel (PDCCH), which carries downlink control information (DCI). The DCI consists of scheduling required for data reception and for the permissions for uplink data transmission by UE
- Physical downlink shared channel (PDSCH), which carries downlink data. The information about where to find this data is defined by DCI carried by PDCCH.
- Physical broadcast channel (PBCH), used for carrying the broadcast information for the UE to use the network.
- Physical uplink control channel (PUCCH), similar to PDCCH, used to carry uplink control information such as the information whether the downlink data transmission is successful or not, a scheduling request (SR) to access the network.
- Physical uplink shared channel (PUSCH), similar to PDSCH, which carries uplink data transmission.
- Physical random access channel (PRACH), the channel used by UE to request connection setup via random access procedure.

Hence the control signaling is carried by PDCCH and PUCCH and data signalling is carried by PDSCH and PUSCH in downlink and uplink respectively. This thesis focuses on reducing latency and improving reliability for PDSCH and PUSCH channels (actual data transmissions). Hence it is assumed that the control information is communicated between base station and UE and perfectly accessible.

3.5 PHY Layer Enhancements for URLLC

URLLC was created by 3GPP to support new services and applications with strict latency and reliability requirements. In this section, a current overview of URLLC with a focus on the difficulties and solutions related to the physical layer in 5G NR downlink and uplink is discussed. The physical layer challenges and enabling technologies, including as frame structure, configured grants in uplink, new modulation and coding scheme tables and adaptive modulation which have been covered in the 3GPP Releases 15 & 16 standardization is presented.

3.5.1 Latency Improvement

Latency is a QoS parameter and it controls how well the targeted communication link performs. Different applications call for various latency levels. As URLLC services demand a very tight latency requirements, 3GPP proposed certain changes in physical layer to achieve these requirements. The following section describes the enhancements in uplink and downlink to support low latency applications.

A. Uplink Configured Grant Transmissions

In uplink, the UE can access the resources when it has data to sent. The UE has to request the base station to allocate the resources to send the data in uplink. In LTE, the uplink service is initiated by the user, and then the enhanced NodeB base station (eNB) initiates a reasonable grant after identifying the uplink service requirement. Such scheduling based transmission also known as dynamic grant (DG) transmissions. Thus the uplink transmission process in LTE is as shown in Fig 3.3 includes:

- SR by UE through PUCCH
- Uplink scheduling by base station to obtain buffer status
- buffer status report (BSR) by UE
- UL scheduling by eNB (enhanced NodeB) for data transmission through PDCCH
- uplink data transmission by UE using PUSCH

This process would take atleast 10 ms before actual data transmission [20], which is not viable for URLLC services. Therefore, a new technology has been introduced in NR which is known as configured grant (CG) based transmissions [11]. With this, a grant-free access in which the network allocates certain resources to UE and UE can send its feedback or the data in buffer in uplink without waiting for a grant also known as Transmission without grant (TWG) to base station to facilitate URLLC services. Hence, the normal handshake delay, such as sending the scheduling

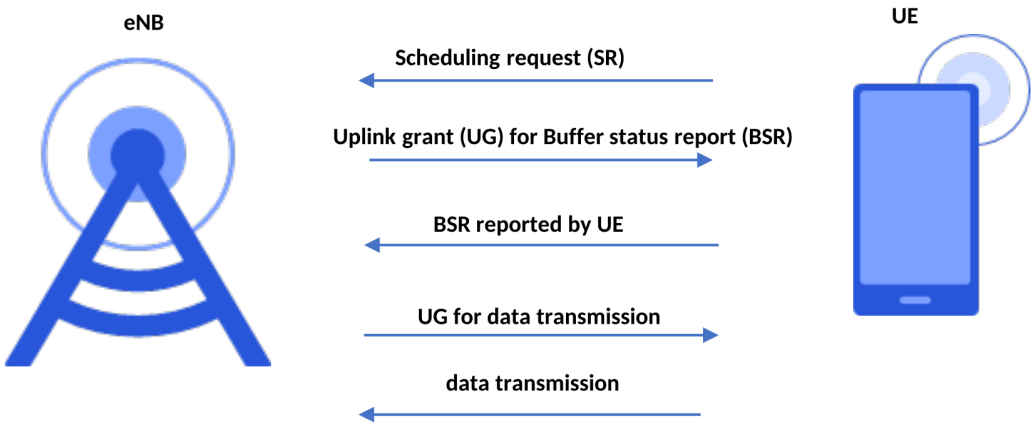

Figure 3.3: Dynamic grant (DG) transmissions in 4G LTE

request and waiting for UL grant allocation, can be avoided by transmitting the data without grant. The basic idea of this technology is that gNB can pre-configured a periodic uplink resources so that whenever uplink traffic happens, a UE can broadcast PUSCH on those resources without any grant, saving time for the handshake reducing initial latency.

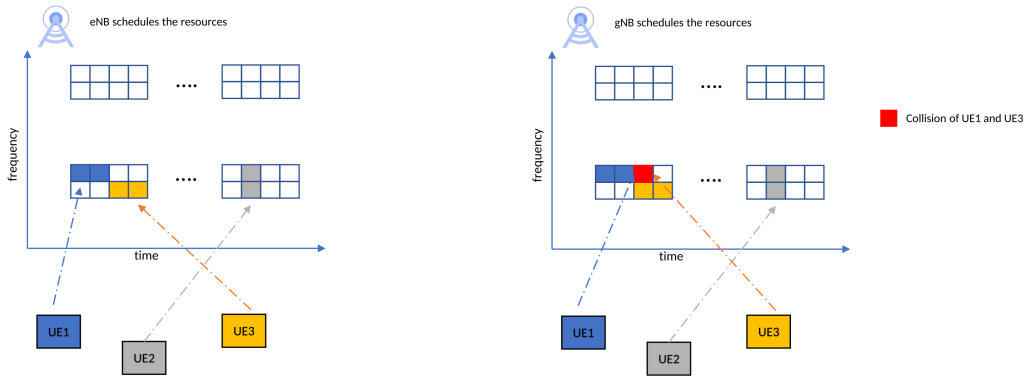

Figure 3.4: DG and CG transmission with collisions in 5G NR

From the latency definition discussed in Section. 2.4.1, the time delay T_s which is responsible for the DG transmission handshake is removed because of CG transmissions. Hence, the physical layer latency is decreased by employing CG in UL data transmissions. Even though the latency is reduced, the challenge that remains is to ensure the reliability of these uplink transmissions. In DG transmissions, the eNB assigns these resources to multiple UEs through DCI which gives information about position of data resources and the UEs perform uplink transmissions using these

resources. Hence, the reliability of data transmissions depends on channel characteristics. However, in case of uplink CG transmissions in multiple user scenario, there is possibility of collisions as there is no scheduled based access for data transmissions as shown in Fig. 3.4. Hence, the reliability of uplink transmission depends not only on channel characteristics but also the collision probability of the UEs resources. Hence the resources should be allocated to URLLC users in uplink based on both collisions and the channel effect.

B. Flexible Numerology and Frequent Transmission Opportunities

As discussed in Section 3.1, the adoption of adjustable sub-carrier spacing is a crucial new feature in 5G. Unlike LTE, where the SCS was fixed at 15 kHz, 5G allows for values of 15 kHz, 30 kHz, 60 kHz, 120 kHz, and 240 kHz in both uplink and downlink [14]. The sub-carrier spacing is inversely proportional to time to transmit the data in the slot. Hence, increasing in sub-carrier spacing decreases the time to transmit the data. Because of this, the duration length of OFDM symbols is reduced, reducing transmission delay T_t from Fig. 2.3. 5G modifies OFDM symbol duration, including CP duration, from a fixed value of 71.35 μs to a set of 71.35, 35.68, 17.84, 8.92 and 4.46 μs utilizing flexibility. As SCS increases, the time to transmit the data (on slot basis) is reduced as shown in Fig. 3.5. The Table. 3.2 shows the different slot duration's for packets with different numerologies specified by 3GPP.

Table 3.2: Different numerology's for 5G

Numerology	**Symbol duration(no CP)(μs)**	SCS(kHz)	**Maximum BW(MHz)**	**slot (ms)**
0	66.7	15	49.5	1
1	33.3	30	99	0.5
2	16.6	60	198	0.25
3	8.33	120	396	0.125
4	4.17	240	397.4	0.0625

Slot-based transmission is utilized in LTE , where one slot represents a transmission time interval. Subslot or mini-slot based transmission is proposed in 5G release to further reduce latency by decreasing transmission time intervals (T_t). A packet is scheduled in a transmission time period of 2, 4, or 7 OFDM symbols. A transmission can begin at the start of the sub-slot transmission time interval, giving it more chances to start in one slot rather than just one in LTE. It shortens the time it takes for an arriving packet in UE buffer to be transmitted thus by reducing T_t.

3.5.2 Reliability Improvement

To decrease the latency, some of the features introduced are discussed in above sections. The 3GPP have specified many features to increase the reliability of data transmissions which includes hybrid automatic repeat request (HARQ), AMC scheme and new modulation coding scheme (MCS) tables.

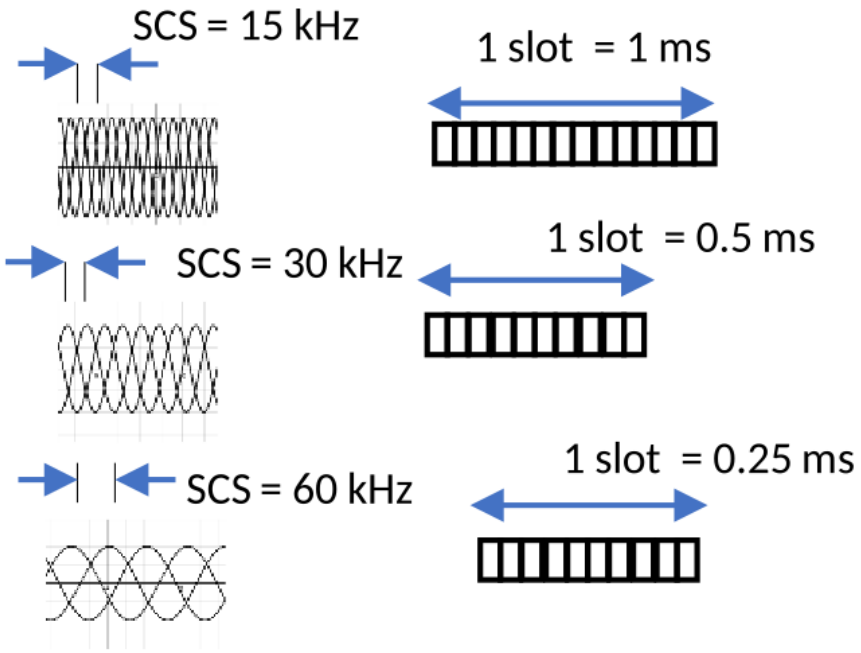

Figure 3.5: Numerologies in 5G NR

A. Hybrid Automatic Repeat reQuest (HARQ) Process

Reliable data transmissions are key for a communication channel for efficient resource allocation and usage. One of the popular techniques used to find errors in communications is channel coding. Redundancy is added to the information through the process of channel coding, which enables the receiver to notice and possibly fix faults [13]. As an illustration, the k information symbols are combined with $n-k$ parity symbols, and the n symbol codeword is sent over a noisy channel. If the receiver can decode the packet, an ACK is sent to sender and if it cannot decode, a NACK is sent by the receiver. To correct these errors, the techniques used are automatic repeat request (ARQ) and forward error correction (FEC). With ARQ, the transmitter sends the undecoded codeword again. With FEC, the receiver may identify and fix faults in the sent bits by using the redundancy that was added to the information. The combination of both error controlling schemes is HARQ. With adaptation of FEC inside ARQ system results in lower amount of retransmissions [13] .

In a wireless radio network, the transmissions from UE to gNB follows the HARQ process [21]. The PHY and MAC layers of 5G NR use the HARQ protocol to provide reliability. HARQ is used to correct erroneous packets sent by the UE. If the data received is incorrect, the receiver (gNB) saves it and asks a retransmission from the sender (UE). After receiving the re-transmitted data, the

receiver uses certain techniques with buffered data before performing channel decoding and error detection [13]. The data in buffer of UE is sent to gNB through HARQ process defined by 3GPP. There are 16 HARQ process IDs in NR [14], whereas in LTE, it is 8.

B. K-repetition CG transmissions

As mentioned, due to CG transmissions in uplink, the latency is reduced. As these resources are not dedicated i.e, not scheduled by gNB, if the network has multiple number of users, these users try to access the resources simultaneously which leads to collision of the data effecting reliability. In order to increase the reliability, 3GPP proposed K-repetition ($repK$) CG HARQ transmissions [11] in uplink. This technique sends a pre-determined number of successive duplicates determined by K of the same packet or transport block (TB) without waiting for input from gNB whether the data is correctly decoded or not as shown in Fig. 3.6, and then the gNB use certain techniques to combine these repeats to improve reliability as has been mentioned before.

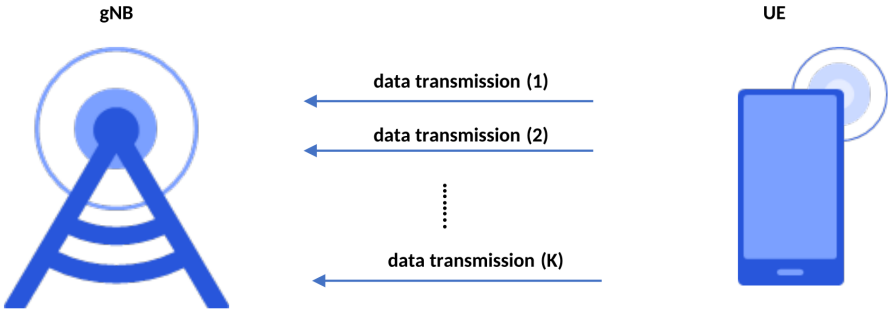

Figure 3.6: K- repetition CG Transmissions

From [22], the block error rate for K transmissions for a HARQ process is given in Eq. 3.4. P_1 is the transmission probability and P_e is the undetected error probability given by Eq. 3.5 [13]. The variables n and $n-k$ represents the message length and length of parity bits respectively. It is observed that the undetected error probability is tends to 0 as the $n >> k$. The value of K is standardized as 1, 2, 4, 8 [23]. The reliability of transmission in uplink is ensured if the UE can perform the repetitions configured by higher layers (RRC)[11].

$$BLER = P_e + P_1P_e + P_1^2P_e + P_1^{K-1}(P_e + P1) = \frac{P_e + (1 - P_e - P_1){P_1}^K}{1 - P_1} \tag{3.4}$$

$$P_e \leq \{1 - (1 - P_r)^k\}2^{-(n-k)} \tag{3.5}$$

An open-source python package called Sionna is used to simulate the block error rate for PUSCH [24]. 3GPP proposed Low density parity check coding (LDPC) and polar coding for data and control channels respectively. The parameters considered for initial simulation are shown in Table. 3.3. The block error rate with respect to SNR with $repK$ repetitions defined in 3GPP is shown in Fig. 3.7. It can be observed by using both HARQ and $repK$ schemes, the reliability of the data transmissions is ensured. Hence, with the K-repetition scheme, the target BLER rate can be achieved. Therefore, as number of repetitions increases, the reliability of the channel increases. If the UE is in cell-edge scenarios, the increase in number of repetitions ensures the reliability.

Table 3.3: Simulation Parameters

Message length (k)	64
Code length (n)	256
Modulation scheme (MCS)	QPSK
Channel	White Gaussian Channel

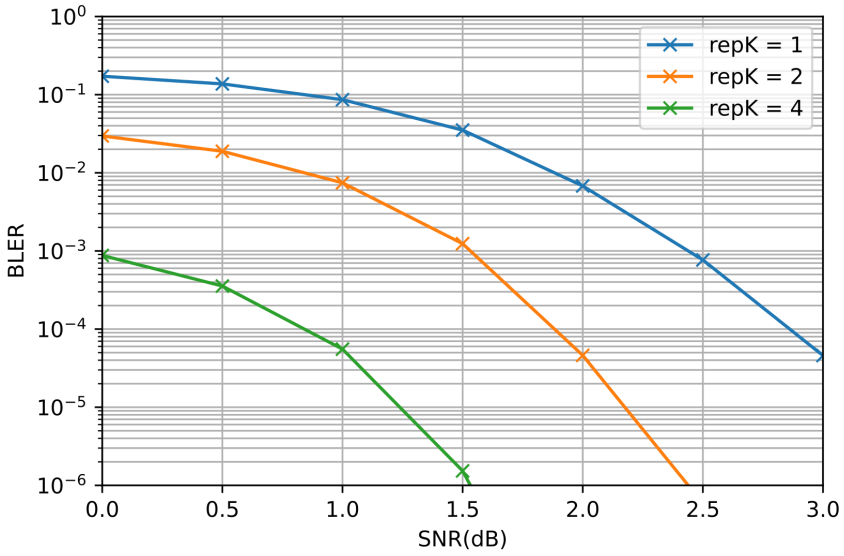

Figure 3.7: BLER with respect to SNR with varying *repK*

C. Adaptive Modulation Coding (AMC) and New Channel Quality Indicator (CQI) Tables for Lower BLER Target

Two key services are anticipated to be supported by the next generation of wireless networks are eMBB and URLLC [25]. Both the services, URLLC and eMBB shares same channel state information (CSI) and MCS tables for target BLER 10^{-1}. The block error rate (BLER) target for the CSI report for URLLC would be 10^{-3} or 10^{-5} than for eMBB (10^{-1}). As a result, in addition to the conventional CQI table with a BLER target which is specified for eMBB, a CQI table matching to this lower BLER target has been established[11]. When choosing a MCS, a lower code rate might increase channel coding redundancy, resulting in greater robustness. As a result, an additional MCS table denoting low target code rates has been developed to correspond to lower BLER target code rates. The optimal code rate and modulation technique for URLLC transmission can be determined using these tables. 3GPP defined the two MCS tables [14]: 64-QAM and 256-QAM. QAM represents a modulation scheme which is quadrature amplitude modulation and the integers represents number of bits per symbol.

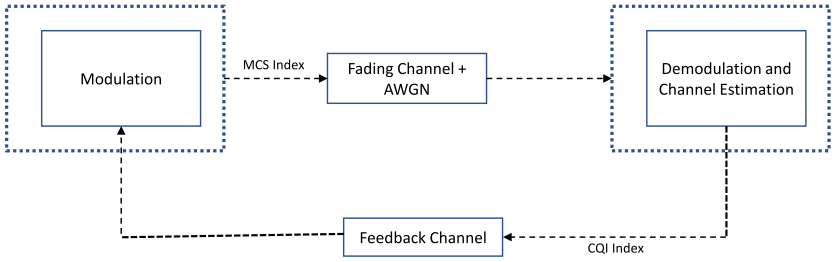

Figure 3.8: Adaptive Modulation Coding based on reported CQI

Based on channel circumstances reported by the UE to the base station, AMC enables selection of the modulation and coding scheme. The physical transmission characteristics can adjust to changes in connection quality thanks to AMC. In general, robust and spectrally efficient transmission over time-varying channels is made possible by adaptive modulation and coding [26]. To change the transmission method in light of the channel characteristics, the primary concept is to estimate the channel at the receiver and relay this estimate back to the transmitter, as shown in Fig. 3.8. When the channel quality is poor, modulation and coding methods that do not adjust to fading conditions fail to provide the required data rates.

As an illustration, in the event that connection quality deteriorates, the PHY layer can fall back to a modulation scheme from the MCS tables that is more noise-resistant. By taking advantage of good channel conditions to send at higher data rates and adapting to the channel fading can increase average throughput, thereby reducing the average probability of bit error. By selecting an appropriate MCS that results in a lower BLER based on reported SNR as shown in Fig. 3.9, it is evident to ensure the reliability of transmissions. It is clearly observed as the reported SNR deteriorates, then a lower MCS value can be chosen for the same reliability target.

3.5.3 Network Slicing

Network slicing enables the division of a physical network into numerous virtual networks (i.e, logical segments) that can support various radio access networks or various types of services for different customer segments, significantly lowering network construction costs through more effective use of communication channels. Network slicing essentially enables the construction of numerous virtual networks on top of a single physical infrastructure [10]. In this virtualized network model, capacity is allocated to specific uses on an as-needed basis. Network slicing enables the construction of slices dedicated to logical, self-contained, and partitioned network tasks. In order to achieve a specific QoS [27] levels defined by standards, such as assured latency, throughput, reliability, and/or priority, network slicing permits the establishment of virtual networks [10] as shown in Fig. 3.10.

Operators can adapt their network for various applications and clients thanks to network slicing. Slices can differ in the services they provide (such as priority and security), the performance they deliver (such as latency, availability, reliability, and data rates). Additionally, this slicing is carried out on both the core network (CN) and the radio access network (RAN) which is a base station [25]. This thesis focuses on RAN slicing while allocating the resources. The 5G system must support

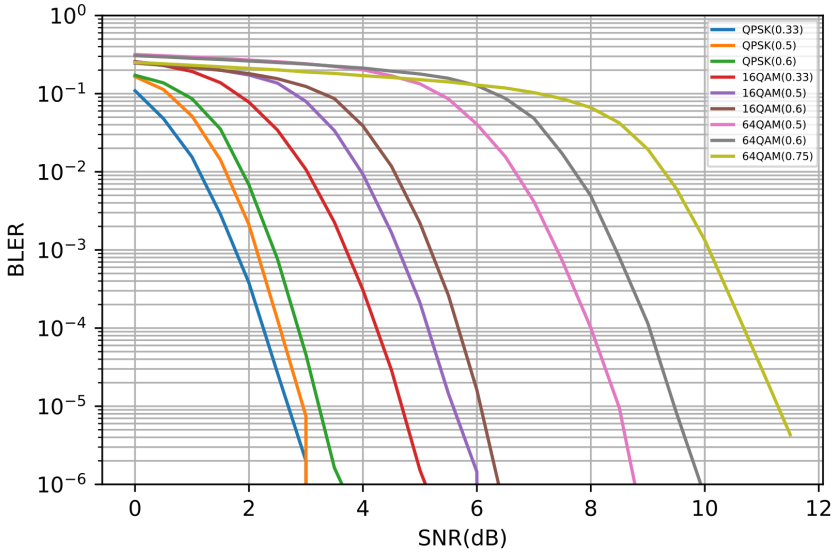

Figure 3.9: BLER with respect to different MCS

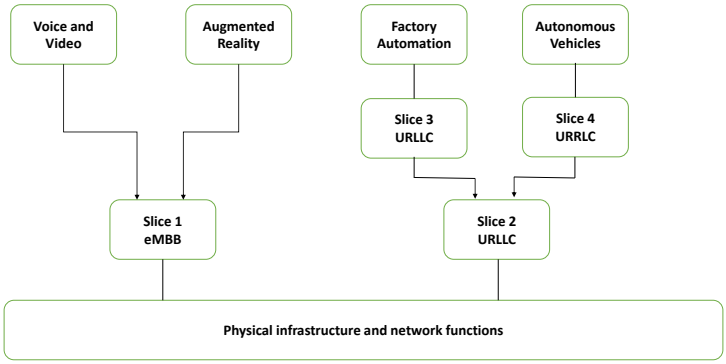

Figure 3.10: Network Slicing for eMBB and URLLC services

the different services. By creating two different slices and allocating the resources independent of each other, the service requirements like latency and reliability can be achieved.

Chapter 4

Downlink Resource Allocation

In this chapter, the resource allocation for mixed traffic composed of two different services, eMBB and URLLC users in the downlink is discussed. When the network has only URLLC users, the resource allocation depends on the tight latency and reliability requirements. While URLLC applications demand these features, eMBB applications demand larger data rates and continuous service. When allocating resources based on service requirements, the fact that the two types of services are multiplexed on the same communications infrastructure provides a challenging resource allocation dilemma.. Hence, the resource allocation problem is considered an assignment optimization problem with data sum rate maximization as an objective. By considering the resource allocation problem as an assignment problem with total data sum rate maximization of resources subject to URLLC latency and reliability criteria, an optimization problem (Appendix B) can be formulated.

4.1 Users in the Network: URLLC

The URLLC service demands tight latency and high reliability. Hence the conventional resource allocation algorithms discussed in Section. 3.3 are not applicable directly as they do not take the QoS requirements of users into account. This section describes the system model and resource allocation optimization problem with latency and reliability constraints when only URLLC users are present in the network.

4.1.1 System Model

Firstly, considering the presence of only URLLC users (U) in the network, the system model is comprised of

- The gNB uses the available radio resources to provide services to all of the URLLC UEs in the cell.
- The available resources are distributed in both the time and frequency domains, as shown in Fig 4.1. Here, the basic sub-carrier spacing of 15 kHz is considered. The same system model can be applied to different numerologies. From this, the time dimension of the resources is divided into N time-slots indexed by $i = 1, 2, ...N$, and the downlink available bandwidth

is divided into F sub-bands indexed by $j = 1, 2, ...F$. The number of time slots in a frame depends on the sub-carrier spacing used.

- A RB, is made up of 12 consecutive sub-carriers and 7 OFDM symbols in a time-slot (0.5 ms).

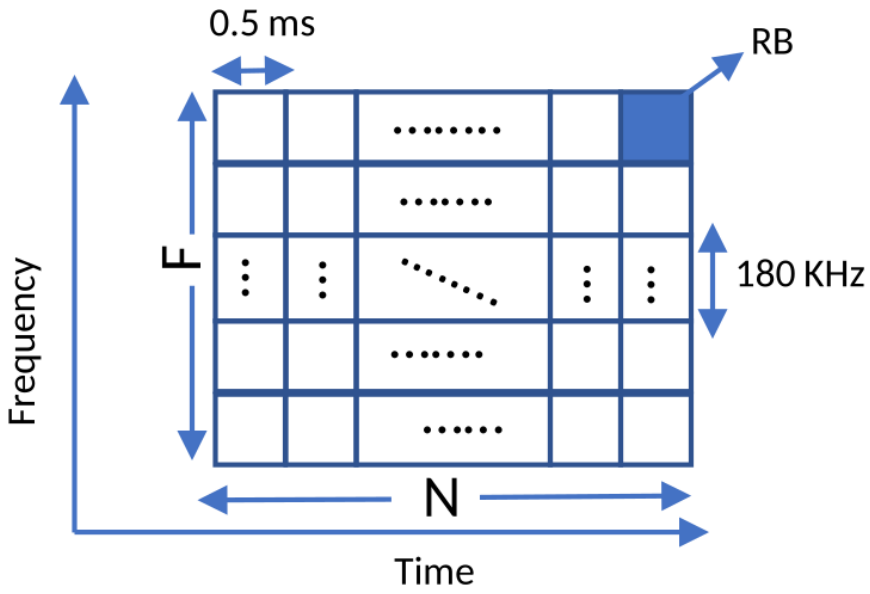

Figure 4.1: Time and frequency radio resource frame

The signal-to-noise ratio (σ) for user e (belongs to URLLC) on sub-band j during time slot i is then calculated as [25]

$$\sigma_{i,j}^{e} = \frac{P|h_{i,j}^{e}|^2 d_{BS,e}^{-\alpha}}{\delta^2} \tag{4.1}$$

where P is the allocated power by base station, $d_{BS,e}$ is the distance between the base station and the UE, α is the path loss exponent, δ^2 is the noise power, and $h_{i,j}^{e}$ represents the channel gain of user e on a sub-band j during the time slot i. It can be observed from the given equation that as the power of the base station increases, the SNR increases. As mentioned in Sec. 3.5.2, increase in SNR leads UE to report higher CQI to the base station, and the base station can assign a higher MCS to achieve a high data rate. The assumption is that the base station's power (P_m) is distributed evenly across all sub-bands. This implies that the power allocated to each band is given by $P = \frac{P_m}{F}$. The CQI depends on many factors that are related to link adaptation techniques. For simplicity, CSI is perfectly accessible at the BS for all users associated with the various services is considered.

The minimum SNR threshold is established in order to attain the right MCS based on the target block error rate and the user-provided channel quality indicator feedback. The MCS table with

different BLER targets for eMBB and URLLC is considered. But the MCS tables specified by 3GPP have only the CQI index, modulation scheme, code rate, and spectral efficiency. However, the data rate is a function of SNR. So, in order to calculate the data rate supported, the MCS table should include the CQI indexes with respect to SNR. One way to have SNR tables is by simulating the cell scenario in Atoll, a commercial software for network planning [28]. Another way to use the coding techniques introduced by 3GPP for 5G NR. The former one with MCS table shown in Table. 4.1 is considered for the simulation [25]. From the following 64 - QAM table, it is evident that the SNR required for a high BLER or reliability target for URLLC is higher than that of the eMBB service for the same spectral efficency.

Table 4.1: 64-QAM MCS Table

Index	**Modulation Scheme**	**Code Rate**	**SNR Threshold [dB] BLER = 0.1**	**SNR Threshold [dB] BLER = 0.001**	**Efficiency [bits/symbol]**
MCS1	QPSK	1/12	-6.5	-2.5	0.15
MCS2	QPSK	1/9	-4.0	0.0	0.23
MCS3	QPSK	1/6	-2.6	1.4	0.38
MCS4	QPSK	1/3	-1.0	3.0	0.60
MCS5	QPSK	1/2	1.0	5.0	0.88
MCS6	QPSK	3/5	3.0	7.0	1.18
MCS7	16QAM	1/3	6.6	10.6	1.48
MCS8	16QAM	1/2	10.0	14	1.91
MCS9	16QAM	3/5	11.4	15.4	2.41
MCS10	64QAM	1/2	11.8	15.8	2.73
MCS11	64QAM	1/2	13.0	-17	3.32
MCS12	64QAM	3/5	13.8	17.8	3.90
MCS13	6AQAM	3/4	15.6	19.6	4.52
MCS14	64QAM	5/6	16.8	20.8	5.12
MCS15	64QAM	11/12	17.6	21.6	5.55

The MCS allows for the following expression of the user e's bit rate when operating in sub-band j during time slot i [25]:

$$R^e_{i,j} = B.T.\mathcal{G}(\sigma^e_{i,j}) \tag{4.2}$$

where $\mathcal{G}(\sigma)$ is the spectral efficiency of the chosen MCS from the MCS table in accordance with the received SNR and B is the bandwidth of an RB, T is the transmission time duration of each slot which depends on sub-carrier spacing. It is clearly evident from the equation that the increase in spectral efficiency with respect to SNR results in an increase in the data rate. The values of B and T are depends on sub-carrier spacing chosen. The data transmissions happen when the RB is allocated to a user. Each RB is given to just one user for the duration of the scheduling round. Hence, the decision variable $s^e_{i,j}$ shows that RB is assigned or not to a URLLC user in the time-frequency resources. Therefore, the binary constraint can be expressed mathematically as:

$$s_{i,j}^{e} = \begin{cases} 1; & \text{if RB is allocated to URLLC user } e \\ 0; & \text{if RB is not allocated} \end{cases} \tag{4.3}$$

From Eq. 4.3 and Eq. 4.2, the sum of bit rate of all users belongs to URLLC service, D_u is given by

$$\mathcal{D}_u = \sum_{e=1}^{U} \sum_{i=1}^{N} \sum_{j=1}^{F} s_{i,j}^{e} R_{i,j}^{e} \tag{4.4}$$

As a result, the resource allocation problem can be transformed into an optimization problem in order to maximize the data rate in the downlink while ensuring latency and reliability. To assign the frequency resources, an orthogonality constraint that ensures that an RB can only be assigned to one user among multiple users within a time window can be expressed mathematically as [25]:

$$\sum_{e=1}^{U} s_{i,j}^{e} = 1; \forall i, j \tag{4.5}$$

Based on received CQI, the BS assigns MCS for the target BLER for URLLC service. From Table 4.1, the BS can choose respective MCS based on SNR received with target BLER. Here, a BLER target of 10^{-3} is used for URLLC, but the same study can be applicable to higher BLER targets (10^{-5}). Therefore, the reliability constraint is satisfied through adaptive modulation scheme. The Eq. 4.6 is considered as latency constraint to ensure URLLC consumers latency requirements are met. To maintain the user's required latency, the URLLC user must get at least one RB for every N time slots when the URLLC user is scheduled [25].

$$\sum_{j=1}^{F} \sum_{i=kN+1}^{kN+N} s_{i,j}^{e} >= 1; k = 0, 1, 2, ...; e \in U \tag{4.6}$$

It is important to ensure the assigned resource block to a URLLC user must transmit the certain size of the packet. Hence, from [29], it is considered that each of the URLLC user's allotted RBs must at least transmit a whole data packet . In this aspect, Eq. 4.7 imposes the requirement that each planned RB for the URLLC user transmits more data than the required threshold (i.e., one packet size denotes by R_m measured in bits). It signifies that the rate produced by an allocated RB should be larger than or equal to R_m amount of bits when $e \in U$. As a result, this constraint makes it possible to send at least one packet with required size through the designated RB (when $s_{i,j}^{e} = 1$).

$$R_{i,j}^{e} \geq R_m; \forall e \in U \tag{4.7}$$

However, the constraint (Eq. 4.7) depends on power (P_m) of base station. The power is assumed to be equally divided among all RBs. The above constraint forces the RB to have a data rate (in bits) to be more than R_m. This is not the case when the power of base station is less. There exist a optimal solution to the problem subjected to this constraint when the power of base station is sufficient enough to have $R_{i,j}^{e}$ more than threshold. The latency constraint implies that atlease one RB allocation for each frame. Hence, to increase the frequency diversity for URLLC user and to send the whole packet, the following constraint is introduced:

$$\sum_{i=1}^{N}\sum_{j=1}^{F} s_{i,j}^{e} \cdot R_{i,j}^{e} \geq R_m; \forall e \in U \tag{4.8}$$

Hence, the optimization problem can be reformulated based on latency and reliability constraints as follows:

$$\begin{aligned} \max_{s_{i,j}^{e}} \quad & \mathcal{D}_u \\ \text{s.t.} \quad & s_{i,j}^{e} \in \{0,1\}; \\ & \sum_{e=1}^{N_U} s_{i,j}^{e} = 1; \forall i,j \\ & \sum_{j=1}^{F}\sum_{i=kN+1}^{kN+N} s_{i,j}^{e} >= 1; \; k = 0,1,2,...; \forall e \in U \\ & R_{i,j}^{e} \geq R_m; \forall e \in U \; or \\ & \sum_{i=1}^{N}\sum_{j=1}^{F} s_{i,j}^{e} \cdot R_{i,j}^{e} \geq R_m; \forall e \in U \end{aligned} \tag{4.9}$$

In optimization theory, the problem is solvable when the objective function is differentiable and constraints should be inequalities in the form of convex problems. But the objective function has a spectral efficiency parameter which is a step-wise function, thus making the optimization problem difficult to solve analytically. In order to solve this, from [25] [13], an approximation function based on received SNR and target BLER for URLLC services are considered as

$$\mathcal{G}_U(\sigma_{i,j}^{e}, \beta_U) = -\log_2(1 + \frac{\sigma_{i,j}^{e}}{\Omega_U}) \tag{4.10}$$

where $\Omega_U = -\frac{\log_e(5\beta_u)}{1.5}$ represent the gaps of SNR, β_U gives the target BLER from MCS table for URLLC service. The data rate values from MCS table are approximately contrasted with the suggested approximations functions shown in Fig 4.2.
One more problem to solve is combinatorial nature of objective function (Eq. 4.3). Because of this, searching through all potential users and subcarrier allocations is necessary to discover the best solution to this non-convex problem, which is prohibitively difficult to use in multi-user resource allocation. Hence, the constraint on the assignment variable $s_{i,j}^{e}$ is relaxed to $0 \leq s_{i,j}^{e} \leq 1$ using time sharing condition, which permits time sharing of each subcarrier in order to avoid combinatorial character [30] [31]. This integer constraint can be loosened to a continuous one, allowing each subcarrier to time sharing [32].

It was proven that when the number of subcarriers increases to higher values, the time-sharing condition under which the duality gap is zero is always satisfied in OFDM systems [32]. The cause is that channel conditions in adjacent subcarriers are frequently identical in real-world OFDM systems with several subcarriers [30]. Following that, each subcarriers time sharing might be roughly implemented along with the frequency sharing of these nearby subcarriers. Now, the objective function is concave and the constraints are convex, therefore an optimal solution exists with local minimum or maximum as global optimum solution. The formulated optimization problem can be

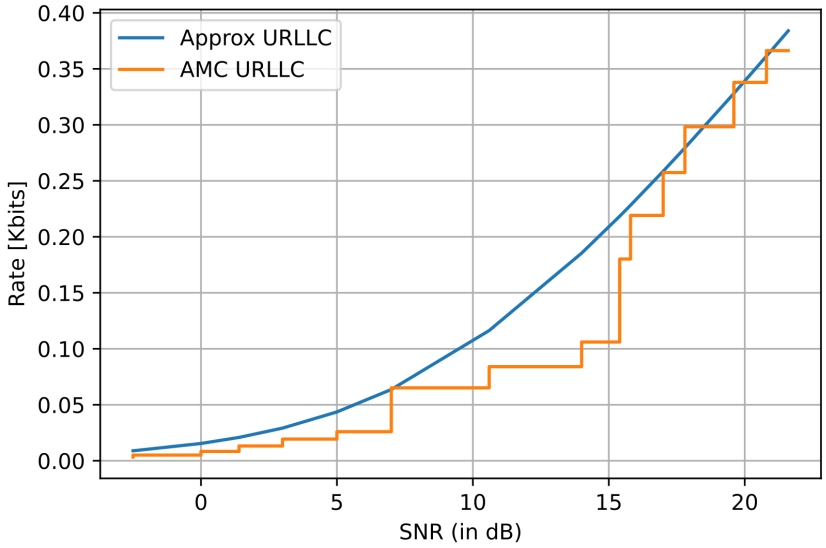

Figure 4.2: URLLC services data rate using MCS table and approximated functions

solved using the common available tools (Guropy [33], CVXPY ..) in python to find the optimal points. Now, the objective function to maximize sum-rate of all URLLC users in the network can be rewritten as

$$
\begin{aligned}
\max_{s^e_{i,j}} \quad & \mathcal{R} \\
\text{s.t.} \quad & 0 \leq s^e_{i,j} \leq 1; \\
& \sum_{e=1}^{N_u} s^e_{i,j} \leq 1, \forall i, j \\
& \sum_{j=1}^{F} \sum_{i=kN+1}^{kN+N} s^e_{i,j} >= 1;\ k = 0, 1, 2, ...;\ \forall e \in U \\
& R^e_{i,j} \geq R_m;\ \forall e \in U \ or \\
& \sum_{i=1}^{N} \sum_{j=1}^{F} s^e_{i,j} \cdot R^e_{i,j} \geq\ R_m;\ \forall e \in U
\end{aligned}
\tag{4.11}
$$

The suggested technique simultaneously optimises the resource allocation across frequency and time, encompassing the entire frame. The final output indicates which consumers should be serviced on which RB in each scheduling round.

4.2 Users in the Network: eMBB + URLLC

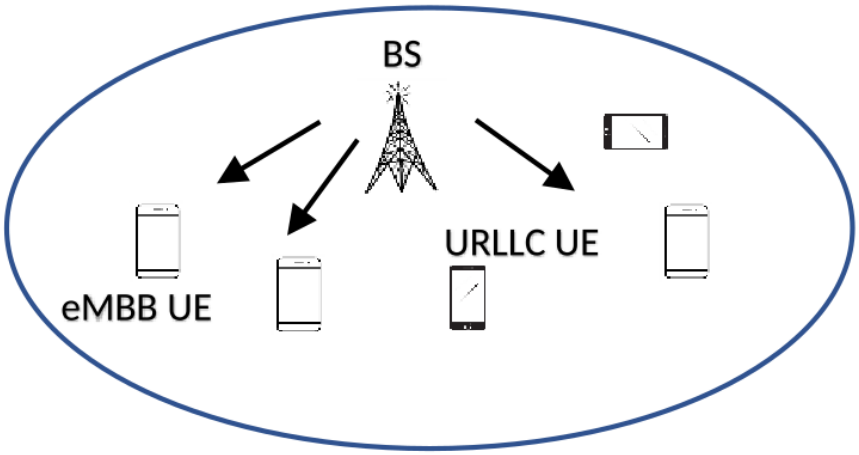

Figure 4.3: Cell network with both eMBB and URLLC UEs

A network scenario for a single-cell wireless cellular network, where UEs are dispersed uniformly over the network and a base station is situated in the cell's centre as illustrated in Fig. 4.3 is considered for the different services simulation. These UEs distributed over the network are connected to various services, such as eMBB and URLLC, indicating that there are several services available in the network. The number of users are divided into E eMBB and U URLLC users. Likewise derived for URLLC, the sum of bit rate of all eMBB users, $\mathcal{D}_e$ is given by [25]

$$\mathcal{D}_e = \sum_{e=1}^{E} \sum_{i=1}^{N} \sum_{j=1}^{F} s_{i,j}^{e} R_{i,j}^{e} \tag{4.12}$$

Following the same approximation for spectral efficiency in URLLC, a differentiable function which makes the objective function easily solvable is [25]

$$\mathcal{G}_E(\sigma_{i,j}^{e}, \beta_E) = -\log_2(1 + \frac{\sigma_{i,j}^{e}}{\Omega_E}) \tag{4.13}$$

where $\Omega_E = -\frac{\log_e(5\beta_E)}{1.5}$ represent the gaps of SNR, β_E gives the target BLER from MCS table for eMBB service. The data rate values from AMC table are approximately matched with approximation functions plotted for eMBB shown in Fig. 4.4. Then the objective function which is the sum of bit rate of users (both eMBB and URLLC) in network is given by

$$\mathcal{R} = \mathcal{D}_e + \mathcal{D}_u \tag{4.14}$$

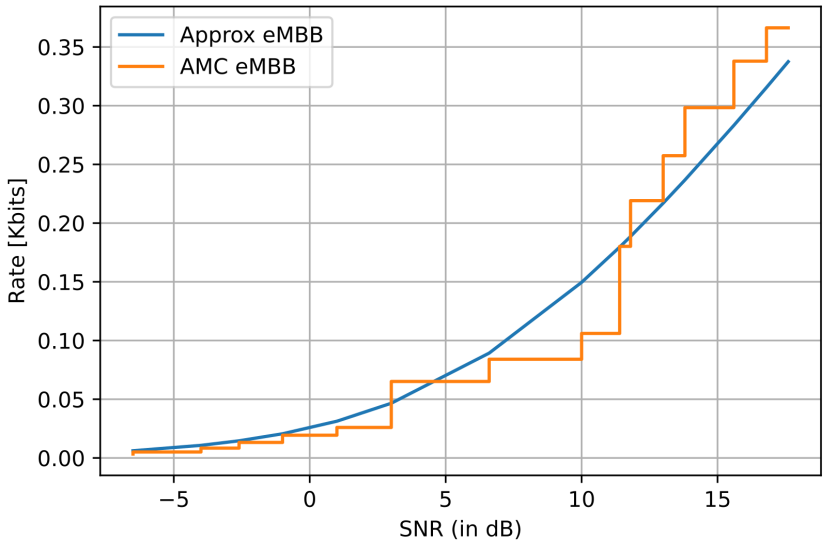

Figure 4.4: eMBB services data rate using true MCS and approximated functions

The concept of network slicing is introduced to create a slices of radio network with respect to quality of service requirements. The system model considered has to provide two different services: eMBB service which demands higher throughput and BLER target 0.1 and URLLC service which demands low latency and high reliability target. Hence, from [25], an orthogonality constraint which ensures that an RB can only be assigned to one user who is a member of one service within a time window can be expressed mathematically as:

$$\sum_{e=1}^{E}\sum_{e=1}^{U} s_{i,j}^{e} = 1; \forall i, j \tag{4.15}$$

Finally, Eq. 4.16 is utilized to ensure a minimum rate R_{min} and validates that each scheduled eMBB user transmits at least R_{min} bits each frame to maintain continuous service [34] [29].

$$\sum_{j=1}^{F}\sum_{i=1}^{N} s_{i,j}^{e} \cdot R_{i,j}^{e} \geq R_{min}; \forall e \in E \tag{4.16}$$

The main goal of the suggested optimization problem is to maximise the overall sum-rate of users associated with eMBB and URLLC services by allocating dynamic resource blocks while taking into consideration a number of restrictions. Hence, the optimization problem for which it is to be maximized is defined mathematically as

$$
\begin{aligned}
\max_{s^e_{i,j}} \quad & \mathcal{R} \\
\text{s.t.} \quad & 0 \leq s^e_{i,j} \leq 1; \\
& \sum_{e=1}^{E} \sum_{e=1}^{U} s^e_{i,j} <= 1; \forall i, j \\
& \sum_{j=1}^{F} \sum_{i=kN+1}^{kN+N} s^e_{i,j} >= 1;\ k = 0, 1, 2, ...;\ e \in U; \\
& \sum_{j=1}^{F} \sum_{i=1}^{N} s^e_{i,j} \cdot R^e_{i,j} \geq R_{min}; \forall e \in E; \\
& R^e_{i,j} \geq R_m; \forall e \in U \ or \\
& \sum_{i=1}^{N} \sum_{j=1}^{F} s^e_{i,j} \cdot R^e_{i,j} \geq R_m; \forall e \in U
\end{aligned}
\tag{4.17}
$$

The algorithm to assign the resources as follows:

Algorithm 1 Downlink Resource Allocation Algorithm

procedure RB ASSIGNMENT(B, U, E, N, T, F, R_{min}, R_m, P_m) ▷ Inputs
  for each all users in the network $e = 1 : U + E$ **do**
    for each time slot $i = 1 : T$ **do**
      for each resource $j = 1 : N$ **do**
        Do this
        Calculate $h^u_{i,j}$ (based on channel)
        Calculate $\sigma^e_{i,j}$ using Eq.4.1
        Find data rate of user $r^e_{i,j}$ using Eq. 4.2
        Assign resources satisfying the Eq. 4.15
        if $e \in U$ **then**
          Assign resources based on constraint Eq. 4.6 and 4.7
        else if $e \in E$ **then**
          Assign resources based on constraint Eq. 4.16
          Calculate $\mathcal{R}$
          Data sum rate of all users = $\arg\max_{s^e_{i,j}} \mathcal{R}$
        end if
      end for
    end for
  end for
end procedure

4.3 Numerical Results

4.3.1 Simulation scenario

A gNB with a 300m radius is placed in the middle of the cell service area. Within the cell coverage region, E eMBB users and U URLLC users are allocated at random. Additionally, a Rayleigh fading channel (Appendix B) between the base station and the users is considered. The simulations are run for a 10 ms frame , which has a bandwidth of 20 MHz and consists of 100 RBs. Each RB has 12 sub-carriers. Please note that the number of sub-carriers per each RB and frame time remains constant, but the slot duration and bandwidth used depends of numerology used. In simulation, each RB's bandwidth is assumed to be 180 kHz. Additionally, the white Gaussian noise power is assumed as 10^{-11}W. The codes are written in python and can be found in GitHub repository [35]. The simulation parameters are shown as follows:

Table 4.2: Simulation Parameters

BS Power (P_m) [dBm]	10, 20, 30, 40, 50
path loss exponent (α)	3
Channel	Rayleigh fading model
Time domain slots (N)	20
RBs per slot (F)	100
Slot length (T) [ms]	0.5
T_{PF} [ms]	10, 30, 50, 70
Number of OFDM symbols per slot	7
Number of sub-carrier per slot	12
R_{min} [Kbits]	30
R_m [Bytes]	32,64,128
SCS [kHz]	15
Carrier BW [MHz]	20
Cell radius [m]	300
Number of eMBB users (E)	10,10
Number of URLLC users (U)	10,40

4.3.2 Comparison of Scheduling Algorithms

In terms of the achieved delivered data rate for eMBB users, a comparison of the performance of the optimized method discussed in this section with that of the baseline methods discussed in above sections is simulated. The empirical cumulative distribution function (ECDF) of achieved delivered data rates of eMBB users on one frame using the suggested methodologies, EDS, RR and

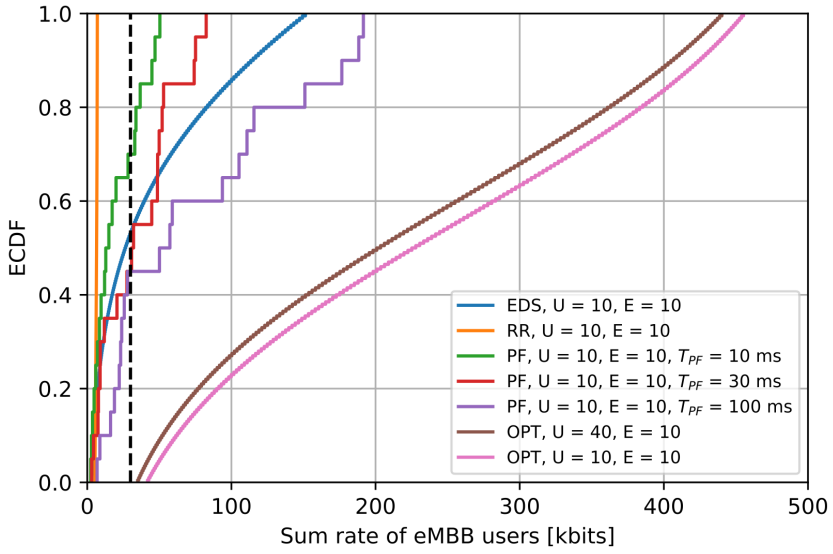

Figure 4.5: Cumulative distribution of eMBB rates (per one frame) using different scheduling algorithms

PF is shown in Fig. 4.5. It is observed that when the network is configured with both eMBB and URLLC users, the RR scheduler is failed to achieve the certain throughput (R_{min}) requirement for eMBB users. It can be observed from the simulation that the RR scheduler provides lower sum throughput for eMBB users. This is because it assigns the resources without depending on channel feedback but assigning resources to all users equally irrespective of QoS. The RR scheduler assigns resources to both eMBB users and URLLC users but failed to achieve the minimum requirements of eMBB users if the sum rate is higher. The EDS is also failed to provide the data sum rate less than 30 Kilo bits (Kbits) (shown as dotted straight line) to 50% of eMBB users in the network. The total number of available resources are divided equally for both eMBB and URLLC users and allocated. As half of the resources are allocated to URLLC, the remaining are allocated to eMBB. These resources are not sufficient to provide the minimum throughput to some of the eMBB users.

The PF scheduler is simulated with various T_{PF} intervals. From [19], it is shown that the throughput increases with increasing time interval. The simulation also shows with increasing in T_{PF} increases the throughput of PF scheduler. It provides lower sum rate than EDS scheduler for certain values of time interval T_{PF}. And, it also fails for some of the users in the network if the minimum sum rate is 30 Kbits. However, by respecting the isolation between service slices, the optimization-based maximum rate scheduling algorithm, provides users with resources and satisfies the minimum rate constraint for eMBB users. In parallel, it also provides the URLLC users with strict reliable and latency requirements. The results show that, the base-line scheduling algorithms does not meet the isolation and minimum rate requirements when compared with discussed optimization technique.

4.3.3 Data Rate of eMBB and URLLC Users

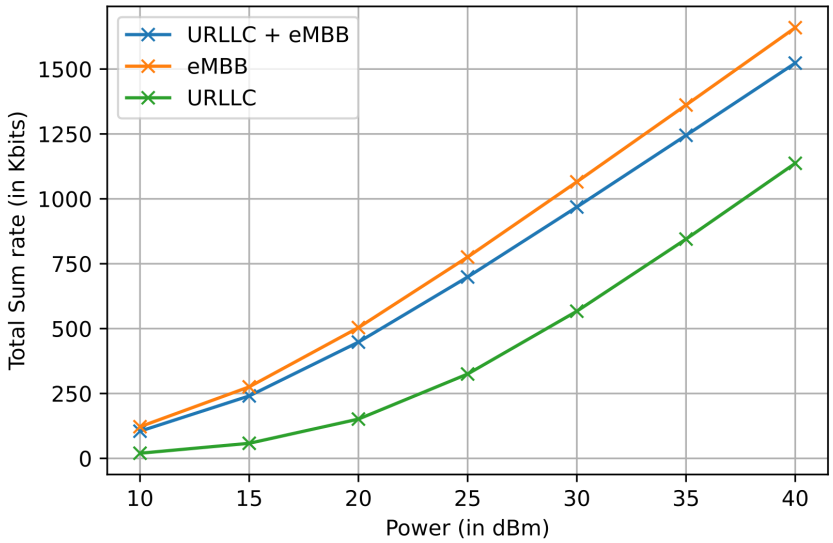

Figure 4.6: Sum rate of eMBB and URLLC users with varying SNR

As expected, it can be shown from the findings that the sum-rate of both eMBB and URLLC users rises with SNR as shown in Fig. 4.6. The reason for this is because when the SNR gets better, the UEs can choose higher order MCS that raises the user's sum-rate. It is observed that when all of the active users in the cell are connected to the eMBB services, the resulting sum-rate of users serves is higher than the cell with only URLLC. As the reliability constraint for eMBB is less, the users choose higher order MCS and result in high data rate. As shown in above, when all the active users in the cell are URLLC, the result of the sum-rate is lower than eMBB service. This is because the URLLC reliability constraint must be satisfied, the users choose lower MCS compared to eMBB.

When the active users in the cell are associated with the two types of services (both eMBB and URLLC), the resulting total sum-rate of the users is higher than the total sum-rate of all URLLC users and lower than the total sum-rate of all eMBB users. This is because that the eMBB users are scheduled with higher MCS and URLLC users are scheduled with lower MCS with respect to SNR. This results in decrease in data rate compared to eMBB but increase compared to URLLC traffic. Hence, it is observed that the optimization algorithm assigns the resources to both eMBB and URLLC users with different reliability targets and 1 ms URLLC latency. Increasing in number of resources assigned to URLLC users may impact on eMBB data rate which is studied further.

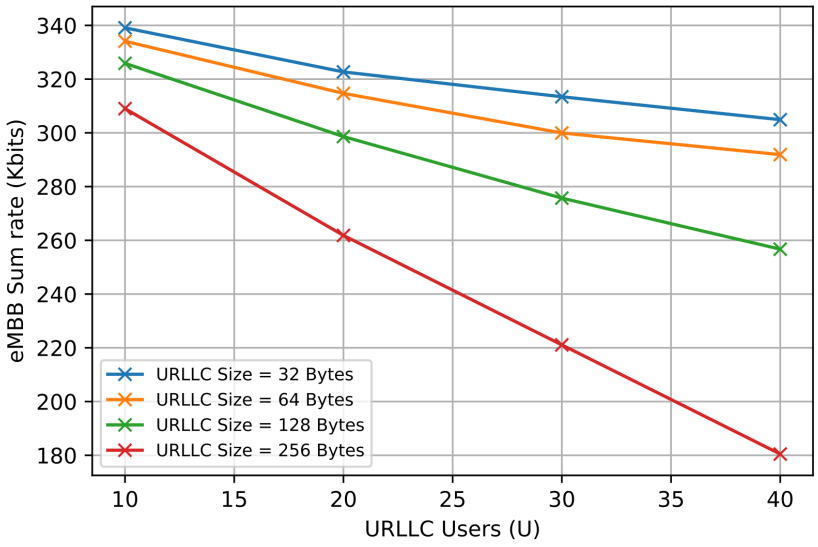

Figure 4.7: eMBB data sum rate for different URLLC users (U) and URLLC Size (R_m) using proposed scheduling algorithm

Fig. 4.7 plots the data sum rate of eMBB users with changing URLLC user and URLLC packet size to be sent over RB's. This study is performed to check the effect of URLLC users and the requirements on eMBB data rate. It is observed that the data sum rate of eMBB users decreases with both URLLC users and packet size. This is because, from Eq. 4.8, as packet size increases, the number of resources in certain latency allocated to URLLC users increases. As number of URLLC users increase, this also results in decrease in eMBB rate. The remaining resources are allocated to eMBB users. Hence, increase in URLLC users and the data requirements, the eMBB users suffer to get enough resources to maintain minimum throughput for continuous service.

Chapter 5

Uplink Resource Allocation

This chapter discusses the resource allocation for URLLC users in uplink presented in the network, focusing on the 3GPP Release 15 feature: configured grant or grant-free transmissions. The resource allocation in downlink is discussed in the previous chapter. To satisfy the reliability and latency requirements of URLLC users in downlink, the AMC scheme and an optimization problem with a latency constraint are considered and solved. However, in uplink, the UE needs permission from gNB to access the resources, which causes latency to increase. Without needing to wait for an uplink permission or resource allocation from gNB, UE can transmit uplink traffic grant-free in order to meet the latency criteria for first access as discussed in Section. 3.5.1. This chapter gives an overview of the Poisson process, contention-based transmission, and optimal resource allocation for URLLC users. A comparison between 3GPP releases and optimal resource allocation for URLLC is presented. Finally, the resource allocation for eMBB users in the presence of URLLC users is discussed.

5.1 Poisson Process

Poisson process is considered one of the most popular counting methods in statics [36]. It is typically employed in situations where the occurrences of specific events seem to occur at a certain rate but are actually completely random. In probability theory, it is a stochastic process that counts the number of events over a certain period of time. Each of these inter-arrival times is supposed to be independent of other inter-arrival times, and each pair of consecutive events has an exponential distribution with an arrival rate parameter [36]. The Poisson process or its expansions have been applied in practice to model the number of users in a wireless network. As per 3GPP [37], the packet arrival follows the Poisson process. Hence, the Poisson arrival process is considered in the derivation of equations to show the relationship between the number of resources and URLLC users in the network. Assuming a random variable τ as the arrival time of an event that is exponentially-distributed. To determine an event will occur in a time interval [0,t], the random variable domains is

$$\begin{cases} 0 < \tau \leq t; & \text{if event happened in time interval} \\ \tau \geq t & \text{if event did not happen in time interval} \end{cases} \tag{5.1}$$

Let the variable λ is the average number of events expected to occur in any given time interval $[t, t+1]$. Then the exponentially distributed random variable's probability density function is represented by the equation when $t \geq 0$ [38]:

$$g(t) = \lambda e^{-\lambda t} \tag{5.2}$$

From Equation 5.1 and 5.2, the probability that the event will take place between $[0, t]$ is

$$P[0 < \tau \leq t] = \int_0^t g(x)\delta x = \lambda \int_0^t e^{-\lambda x}\delta x \tag{5.3}$$

By solving the above equation, the probability that the event will take place between $[0, t]$ is

$$P[0 < \tau \leq t] = 1 - e^{-\lambda t} \tag{5.4}$$

5.2 Users in the Network : URLLC

To reduce latency, uplink transmissions for configured grant-based transmissions access the resources without any information from gNB, as discussed in Section 3.5.1. This is known as contention-based access. When a specific amount of resources is allotted for URLLC services, UEs can access these resources for data transmissions without any DCI. These resources, configured by the network, can be shared among UEs thereby increasing resource efficiency. This may cause packet collisions, and the reliability and performance may degrade as a result. It results in a high error rate. Hence, the data or packet has to be retransmitted again, which may increase the latency. The following sections explain the system model with only URLLC users and derive the relationship between the number of URLLC users and resources by using the Poisson process [39].

5.2.1 System Model: Contention-Based Access

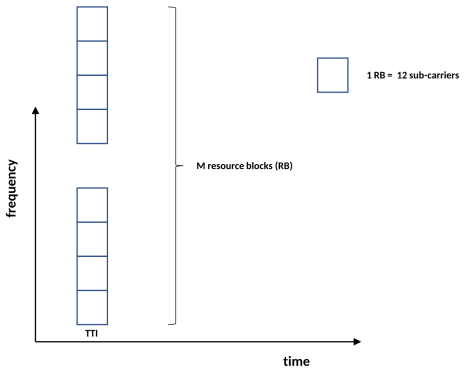

Figure 5.1: M resource blocks shared by U URLLC UEs

The system in Figure 5.1 is assumed to have U URLLC UEs for calculating the probability of data transmission collisions in shared resources. It is also considered that these U UEs are configured

by gNB to transmit the data over shared resources randomly. The packet in each UE follows a Poisson process with exponentially distributed inter-arrival times. The system parameters assumed and URLLC requirements for this model are:

- The item μ represents the average packet inter-arrival time, and the item T represents the total transmitted time (TTI).
- The average number of random access events in an interval is denoted as λ given by T/μ [39].
- Each packet may require an access slot for a resource block in the frequency domain, and a unit TTI in the time domain is assumed.
- The URLLC reliability target is $P_{rel} = 0.999$, and the maximum E2E latency allowed is 1 ms.
- The bandwidth of the system is divided into M resource blocks in one TTI as shown in Fig 5.1. Each RB is comprised of 12 REs.

A. Collision Probability Calculation

If the U UEs in the system need to do transmissions without any grant, they can use any block in the reserved resources of a single transmission with random access. From Section 5.1, the probability of one UE to have one or more random transmission event x in a TTI is [39]

$$P_{ra} = P(x > 0) = 1 - e^{-\lambda} \tag{5.5}$$

The probability that no other UE in the system to have random access event in the same TTI as the UE of interest is

$$P_{U-1} = (1 - P_{ra})^{U-1} \tag{5.6}$$

The probability that u UEs from the set of $U-1$ UEs has random transmission events in the same interval (TTI) is,

$$P_u = \binom{U-1}{u} (P_{ra})^u (1 - P_{ra})^{U-1-u} \tag{5.7}$$

The system model has been assumed to have M resource blocks. Hence, the probability of these u UEs do not have random access event as the UE of interest in the given TTI is [39]

$$P_a(u, 0) = (1 - \frac{1}{M})^u \tag{5.8}$$

These u UEs do not collide with the UE of interest but may collide among themselves. The probability of no collision between any other UE and UE of interest is [39]

$$P_{no_c} = \sum_{u=1}^{U-1} P_u P_a \tag{5.9}$$

Then the final collision probability of the UE of interest in this system is obtained by considering Eq. 5.6 and Eq. 5.9 as follows

$$P_{col} = 1 - P_{no_c} - P_{U-1} \tag{5.10}$$

$$P_{col} = 1 - (1 - P_{ra})^{U-1} - \sum_{u=1}^{U-1} P_u P_a \tag{5.11}$$

$$P_{col} = 1 - \sum_{u=0}^{U-1} \binom{U-1}{u} (P_{ra}(1 - \frac{1}{M}))^u (1 - P_{ra})^{U-1-u} \tag{5.12}$$

By using binominal theorem (Appendix B), the collision probability of the UE of the interest is

$$P_{col} = 1 - (\frac{e^{-\lambda} + M - 1}{M})^{U-1} \tag{5.13}$$

B. Repetitions of the Packet

The probability of a collision can be minimized by using diversity transmission [40]. In this case, a UE sends the same data packet in K subsequent TTIs. In contrast, the resource block used in subsequent TTIs is chosen at random from among the M available blocks. Some packets may collide as a result of this, but others may be successfully received, which reduces the total probability of a collision. Each UE transmits the same data packet several times in K subsequent TTIs with diversity transmissions, as shown in Figure 5.2.

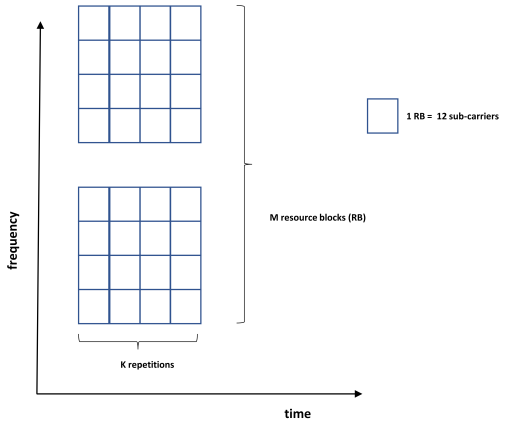

Figure 5.2: Diversity transmission to increase reliability

The M contention resources in TTIs are treated as a new contention pool, from which M^K resources are randomly selected. The traffic intensity for the diversity transmission in the contention pool has been increased to $K\lambda$. The collision probability then becomes [39],

$$P_{cK} = 1 - (\frac{e^{-K\lambda} + M^K - 1}{M^K})^{U-1} \tag{5.14}$$

Hence, based on target collision probability and the number of repetitions of the data packet, the number of resources required for URLLC users is calculated.

5.2.2 3GPP Releases: Solutions and Drawbacks

A. Release 15

With 3GPP Release 15, the UE can broadcast these CG repetitions blindly without receiving feedback from the gNB [41] [23]. The gNB configures the UEs with high priorities and tight requirements to transmit in the GF regions in uplink transmission, as specified. It also configures the K number of times the UE should transmit by RRC layer via the $repK$ parameter to ensure the reliability of the transmission. The values of K are standardized in [23]. It has 1, 2, 4, and 8. Despite the UL grant from gNB, the UE transmits these numbers of repetitions in the allocated configured regions. As mentioned earlier, the uplink data transmissions happen with the HARQ process. Each HARQ process has a unique identification (ID) for the gNB to decode correctly [41]. In order to avoid confusion between different HARQ IDs, UEs are not allowed to transmit these repetitions outside the interval with periodicity defined in the standard. As a result, depending on the arrival of data, the UE must transmit fewer repetitions than configured.

Fig. 5.3 depicts a case in which the number of required repetitions is not guaranteed due to a period P boundary limitation. Assuming the sub-carrier spacing is 60 kHz and the HARQ interval spans the whole slot length, there are four slots whose length is 0.25 ms each for data transmissions. Let's assume the gNB configures the UE to transmit four repetitions so that reliability is ensured. Hence, the interval P contains four CG events, and the UE is expected to do four repeats for every packet. If the data arrives in the buffer before all four CG events in the first period, the UE is able to do four repeats for the first packet, as expected. Hence, the data transmission happens within the required latency, and reliability is achieved. However, there are only three CG instances left in the second period when data arrives after the first CG occasion. 3GPP Release 15 specifies that transmissions must wait until the next same HARQ process ID, which has a significant impact on latency [41]. This suggests that the UE can only perform three repeats, which is less than the expected quantity that is configured by gNB. Similarly, the UE can only transmit twice in the fourth period. When a packet arrives after the first CG event in a period, it is clear that the UE transmits the packet with a lesser number than the number determined by $repK$. It reduces the uplink transmission's reliability. For URLLC UEs with a high reliability demand (10^{-5}), the issue becomes even worse.

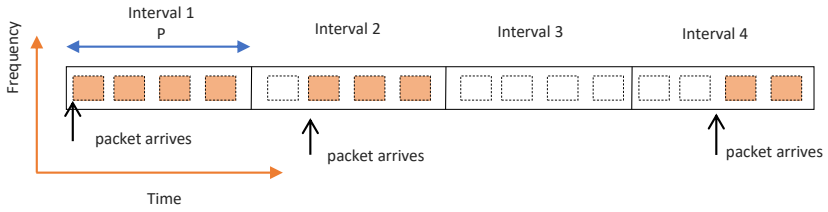

Figure 5.3: Less than K repetitions in CG UL transmission

Furthermore, transmission latency increases as the number of repeats decreases because the gNB is more likely to fail to decode the packet and must plan a retransmission. In that instance, the UE must wait for the gNB to decode the packet repetitions and send a UL permit to reschedule a retransmission if necessary, which has a significant impact on.

B. Release 16: Multiple Configurations

To address Release 15 issue and ensure both repetitions and latency of CG transmissions, 3GPP defined multiple UL CG configurations for each serving cell resource in Release 16 [42]. The idea behind introducing multiple configurations for the applications seen in industrial networks, where multiple data streams generated at a node are frequently used, for example, a robot arm with numerous actuators, sensors, and monitoring devices connected to a single radio module. Thus, as can be seen in Fig. 5.4 , such numerous streams have various characteristics, such as arrival time and payload size.

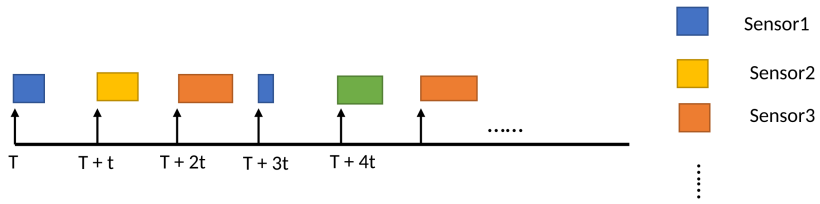

Figure 5.4: Different traffic from sensors in industry with different KPIs

Therefore, even though this configuration allows very short periodicity, these streams cannot be handled by a single GF configuration since they may have different KPIs, such as size, arrival time, MCS index, periodicity, or transmission length. Multiple configurations for a UE within a single serving cell are one possible implementation of CG [43]. With several pre-configured transmission occasions with diverse settings, such as periodicity, time offset, frequency resources, MCS index, etc., the UE is able to send data in a variety of ways. For instance, the blue stream data can be transmitted using configuration 1, the orange stream using configuration 2, and the yellow stream using configuration 3. Configurations 1, 2, and 3 each have a unique offset from the start of the slot or frame. Additionally, resource allocation parameters such as frequency and time duration may vary.

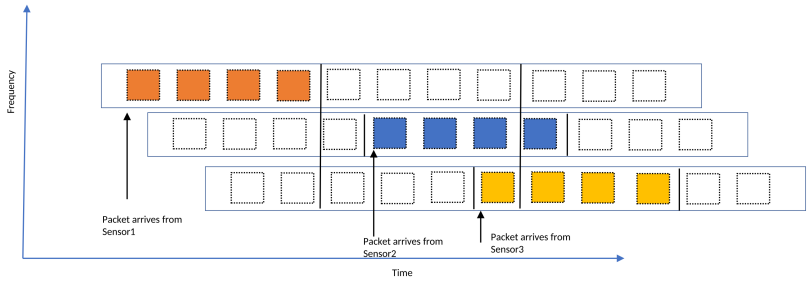

Figure 5.5: Active Multiple Configurations

A UE's CG configurations are limited to 12 per serving cell or base station and are configured by RRC. To ensure that data is always transmitted at the start of a HARQ process interval and that

all configured repetitions are transmitted before reaching the HARQ process boundary, the UE selects the configuration with the earliest starting point to transmit data, as shown in Fig. 5.5 for the case of three active configurations. Hence, by transmitting the configured number of repeats, the multiple configurations ensure both reliability and latency for URLLC users.

5.2.3 Reserved Resources to ensure K repetitions

To ensure that the URLLC UEs meet the stringent requirements, an optimal technique to use the reserved resources to ensure that the UEs can transmit the number of repetitions specified by *repK* at the upper layer is proposed in [41]. The gNB proposes that reserved periodic resources be generated and allocated to a large number of UEs so that they are less likely to retransmit data if transmissions in the reserved resources are required to ensure the specified number of repetitions. The reserved resources are scheduled in the same way as the CG resources. Fig. 5.6 depicts the use of reserved resources.

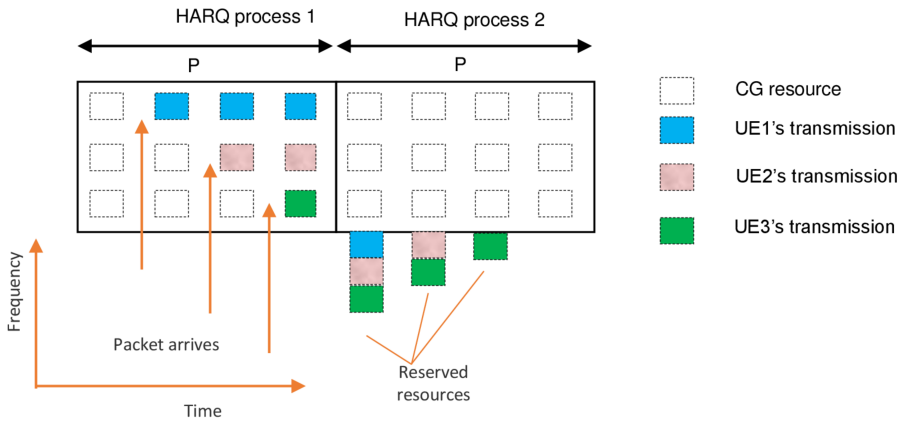

Figure 5.6: Reserved resources for repetitions

Assuming the sub-carrier spacing is 60 kHz and each HARQ process has 4 slots to do the repeats, the gNB schedules the CG resources to three UEs, and each UE is expected to transmit four repetitions of TB. The data for the UE1 comes after the first CG event in the first period; therefore, it can only do three repetitions in that time. To achieve four repeats, the UE1 retransmits data in the first reserved resource of the next period. UE3's data arrives at the last CG event, and only one repetition is possible in CG resources. As a result, the UE3 will use the three reserved resources in the next period to complete the set number of repeats.

These reserved resources are also assumed to be shared among the UEs like configured grant resources, to improve resource consumption efficiency [41]. The first reserved resource in Fig. 5.6 , which has three blocks, is likely to be shared by more than three UEs while still achieving the

target collision probability, which is close to the collision probability of the CG resources. A group of more than two UEs can additionally share the second and third reserved resources. The following sections derive the equation following the same process in contention-based transmission, showing the relationship between the number of UEs, the size of the reserved resources, and the collision probability.

5.2.4 System Model: Optimal Scheme

U UEs are presented in the system depicted in Fig. 5.7 that is used to calculate collision probability in reserved resources. The gNB configures these U UEs to transmit in periodic GF resources with the number of repetitions K determined by the parameter $repK$. The UEs are scheduled to transmit the automatic repetitions of a packet in the consecutive CG resources. The GF resources in a single frequency band can be shared by a group of UEs to improve resource efficiency. The number of GF transmission occasions in a period P corresponding to a HARQ process is equal to the number of repetitions K specified by $repK$ from higher layer.

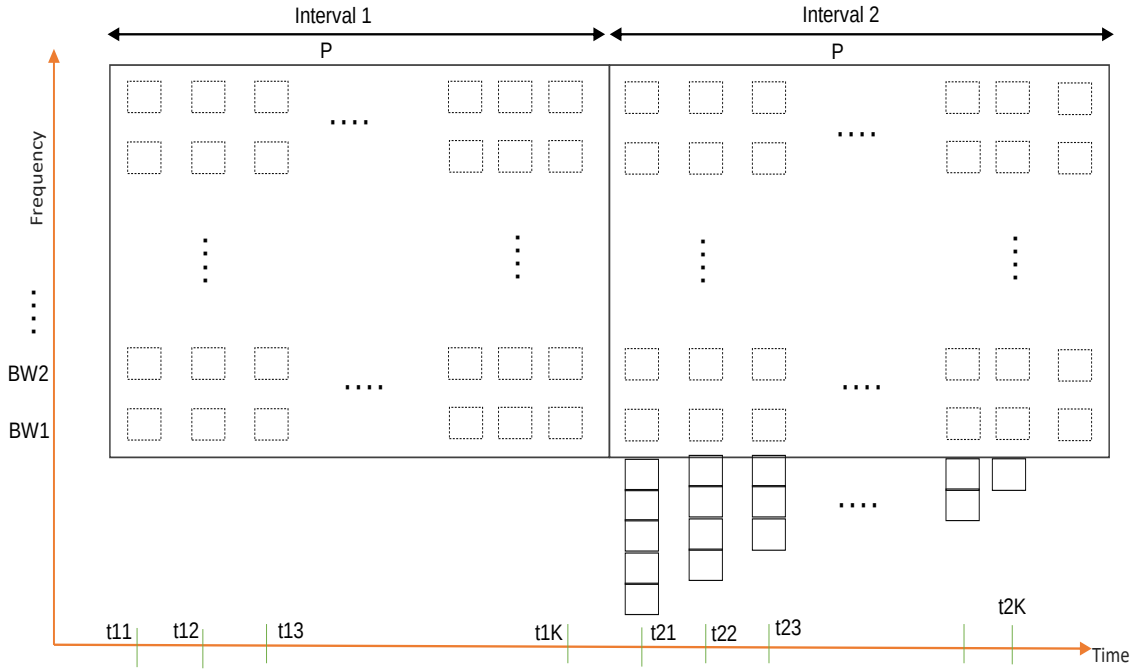

Figure 5.7: System Model

In each period P, the reserved resources are configured as CG resources by gNB. If the U UEs in the system need to make transmissions in order to complete the configured number of repeats, they can use any block in the reserved resources of a certain transmission occasion with random access. Each reserved resource has a block size of M_l, with index l indicating that the reserved resource is at the period's lth transmission event. Each block in the reserved resources has the same size as the GF resource in a single transmission instance.

Collision Probability Calculations

Following the same procedure in contention-based transmission scheme, the collision probability in the reserved resource at the first transmission occurrence of a period (at t21 in Fig. 5.7) is calculated same as before as follows: 1 UE of interest with a transmission in the reserved resource at t21 and the rest of $U-1$ UEs are considered. The probability that one UE has one or more random transmissions from the Poisson process, in time T between two CG resources is

$$P_{data} = 1 - e^{-\lambda} \tag{5.15}$$

Because CG resources in a single bandwidth can be shared by a group of UEs (U_{gUE}), the probability of a CG resource collision between the UE of interest and other UEs in the group using same frequency band is [41]

$$P_{CG} = 1 - e^{-\lambda(U_{gUE}-1)} \tag{5.16}$$

If the data comes after the first CG transmission occasion at t11, the reserved resource at t21 is used by a UE. The probability that no other UE from the set of $U-1$ UEs has a transmission after the first occasion is calculated as follows:

- The probability of UE not to have random access is given by $1 - P_{data}$
- The probability of UE not to have random access after first CG occasion(from t12 till t1K) is given by $(1 - P_{data})^{K-1}$
- By using above two conditions, the probability that no other UE from the $U-1$ UEs has a transmission after the first CG occassion is given in Eq. 5.17 [41]

$$P_d = (1 - P_{data})(1 - (1 - P_{data}))^{K-1} \tag{5.17}$$

If no UE other than the UE of interest has a transmission after the first CG transmission occurrence at t11, there is no collision in the first reserved resource of a period P at t21. After the first CG occurrence, the probability that no other UE from the collection of $U-1$ UEs has a transmission is computed by

$$P_0 = (1 - P_d)^{U-1} \tag{5.18}$$

If there is a transmission from the set of $U-1$ UEs after the first transmission, the probability that u UEs have such transmission is

$$P_u = \binom{U-1}{u} P_d^u (1 - P_d)^{U-1-u} \tag{5.19}$$

At time t21, the probability that the UE of interest and u other UEs do not access the same resource block in the first reserved resource is [41]

$$P_{a0u} = (1 - \frac{1}{M_1})^u \tag{5.20}$$

The probability that the UE of interest does not collide with any other UE in the first reserved resource at t21 is calculated by

$$P_{sum} = \sum_{u=1}^{U-1} P_u P_{a0u} \tag{5.21}$$

From Eq. 5.18 and Eq. 5.21, the collision probability in the first reserved resource for the UE of interest is computed as

$$P_{c1} = 1 - P_0 - P_{sum} \tag{5.22}$$

By using binominal theorem, the collision probability in the first reserved resource for the UE of interest is

$$P_{c1} = 1 - (\frac{M_1 - e^{-\lambda} + e^{-K\lambda}}{M_1})^{U-1} \tag{5.23}$$

In order to generalize the equation, the collision probability in the second reserved resources is calculated. The calculation process follows the same probabilities that have been derived. The probability that no other UE from the set of $U-1$ UEs has a transmission after the second occasion is

$$P_d = (1 - P_{data})^2 (1 - (1 - P_{data})^{K-2}) \tag{5.24}$$

At time t22, the probability that the UE of interest and u other UEs do not access the same resource block in the second reserved resources is

$$P_{a0n} = (1 - \frac{1}{M_2})^u \tag{5.25}$$

The rest of equations remains the same. Hence, the collision probability in the second reserved resources for the UE of interest is calculated as

$$P_{c2} = 1 - (\frac{M_2 - e^{-2\lambda} + e^{-2\lambda}}{M_2})^{U-1} \tag{5.26}$$

Therefore, a general equation of collision probability for the reserved resource at any transmission occasion in a period can be developed using the same calculating procedure as follows:

$$P_{cl} = 1 - (\frac{M_l - e^{-l\lambda} + e^{-K\lambda}}{M_l})^{U-1} \tag{5.27}$$

where $l \in$ [1,K-1] is index which indicates the position of the reserved resource based on the transmission occasion position in a period. If P_{ci} is set to the same value in all reserved resources, according to Eq. 5.27, the sizes of the reserved resources in an interval drop from the first to the last: $M_1 > M_2 > M3 > ... > M_{K-1}$. This means that the reserved resources' sizes are optimized based on their locations. When compared to using the same size for all reserved resources discussed in contention-based transmission , this optimization reduces resource consumption. Further simulations shows that the resource consumption is minimized using this scheme compared to using multiple configurations proposed in 3GPP Release 16.

5.3 Users in the Network: eMBB and URLLC

The issue of uplink resource allocation in 5G NR networks for mixed traffic, which includes users of eMBB and URLLC devices, is presented in this section. A mathematical model for grant-free services is discussed in the above sections for the K-repetitions HARQ. The optimal scheme with reserved resources increases resource efficiency by decreasing the number of resources allocated to URLLC users. As discussed in downlink resource allocation, the URLLC users in uplink demand strict latency and reliability, and eMBB users demand higher throughput with certain reliability. For the uplink, the resources allocated to eMBB users are grant-based, and URLLC users on a grant-free basis are assumed. The scheduling issue is formulated as an optimization problem aimed at ensuring the key performance indicators for URLLC and maximizing the data rate for eMBB users in an uplink.

5.3.1 System Model

In this section, likewise the downlink, the resource allocation optimization issue is formulated using both the grant-free and grant-based models for a single gNB. The system is comprised of U URLLC devices and E eMBB users with a single base station serving them. The system model consists of both configured grant resources and grant-based resources. The information about these resources, which is configured through the RRC layer, is assumed. Based on the number of users' requirements in the network, the gNB can schedule or allocate grant-free or grant-based resources to respective users with service demand. The frequency grid is divided and composed of F frequency resources and N time slots likewise in the downlink. To ensure the K-repetitions in the uplink, the grant-free resources are paired with reserved resources from the optimal scheme. Hence the number of resources required for URLLC users is calculated by adding the CG resources and reserved resources from grant-based resources. The gNB broadcasts the minimum GF slot locations for all URLLC users at the start of each TTI. At the very least, the slots offered should meet their latency and reliability criteria.

A. URLLC Grant-Free Model

When the GF URLLC devices have a packet to send, they are allowed to send it directly to gNB in the shared resource pool in an arrive-and-go fashion without any SR and UL grant. The reliability of these transmissions in uplink depends on both collisions and channel characteristics. In optimal scheme resource allocation for URLLC users, it was shown that the higher layers ensure the reliability of the URLLC transmissions if the UE does $repK$ repeats of the same information block. The relation between the number of resources required based on collision probability and K repetitions is presented in the above sections. These resources are allocated or reserved for URLLC users to ensure reliability and latency constraints.

eMBB Grant-Based Model

The base station must schedule the grant-based resources to eMBB users and at the same time, provide the URLLC devices with adequate resources to satisfy their latency bounds. By suing the same AMC scheme in downlink resource allocation, the reliability of grant-based eMBB transmissions in uplink is also ensured. Similarly a decision variable (scheduling parameter) is required to show that the resource is allocated to eMBB user to define the data rate equations as

$$s_{i,j}^{e} = \begin{cases} 1; & \text{if RB is allocated to eMBB user } e \\ 0 & \text{otherwise} \end{cases} \tag{5.28}$$

The data sum rate for all eMBB users from Section 4.2 is given as [25]:

$$\mathcal{U}_e = \sum_{e=1}^{E} \sum_{i=1}^{N} \sum_{j=1}^{F} s_{i,j}^{e} R_{i,j}^{e} \tag{5.29}$$

where $R_{i,j}^{e} = B.T.\mathcal{F}_a(\gamma, \beta_a), a \in E$ [29]. The system of equations from Chapter. 4 are considered for uplink sum data rate calculations.

5.3.2 Formulation of Problem

Likewise in downlink, the resource allocation problem in uplink is considered as optimization problem for resource allocation. Firstly,the minimum amount of resources to meet the URLLC users criteria for latency and reliability should be allocated. The optimization issue, which tries to maximize the eMBB users rate while satisfying the the latency constraints needs of the URLLC devices is formulated , based on the system model. The formulated optimization problem is subject to various constraints satisfying both URLLC and eMBB QoS. The constraint for non-multiplexing of data for different QoS requirements is (similar to network slicing in downlink) [29]:

$$\sum_{e=1}^{E} s_{i,j}^{e} = 1; \tag{5.30}$$

Similar to the Eq. 4.16, in order to avoid any scheduled eMBB users from going starved, the Eq. 5.31 will ensure the minimum throughput required [29].

$$\sum_{j=1}^{F} \sum_{i=1}^{N} s_{i,j}^{e} \cdot R_{i,j}^{e} \geq R_{min}; \forall\, u \in E \tag{5.31}$$

Finally the optimization problem for resource allocation in uplink with the above mentioned constraints is written as follows:

$$\begin{aligned} \max_{x_{i,j}^{e}, R, k} \quad & \mathcal{U}_e \\ \text{s.t.} \quad & s_{i,j}^{e} \in \{0, 1\}; \\ & P(latency < 1ms) \leq 10^{-3} \\ & \sum_{e=1}^{E} s_{i,j}^{e} \leq 1; \\ & \sum_{j=1}^{F} \sum_{i=1}^{N} s_{i,j}^{e} \cdot R_{i,j}^{e} \geq R_{min}; \forall u \in \mathcal{U} \end{aligned} \tag{5.32}$$

The optimization problem is a class of nonlinear probability constrained programming problems. Hence, it is not solved using standard optimization techniques[34]. The combinatorial nature of

problem can be solved likewise in downlink by converting into continuous one. Therefore, the issue is broken down into two sub-issues: best scheduling for eMBB users and meeting the requirements of the URLLC device with the fewest resources possible. Since the probability of collision is only impacted by the quantity of the allotted URLLC resources, M, the problem is divided as follows:

- The collision probability constraint is firstly solved for a certain number of resources allotted to URLLC devices, U, and repetitions, K.
- After these resources are allocated to URLLC users, the rest of the frequency resources can be allocated to maximize the sum data rate in uplink for eMBB users, because from the eMBB users' perspective, the least amount of frequency resources to meet the URLLC devices latency constraints are assigned by choosing the minimum M (and the corresponding K) to satisfy the URLLC devices latency requirements. This will then result in maximizing the eMBB users rates.

The algorithm to assign the resources in uplink is as follows:

Algorithm 2 Uplink Resource Allocation Algorithm

procedure RESOURCE ALLOCATION(U, λ, P_{col}, K) ▷ Inputs
    Calculate reserved resources M using Eq. 5.27
    Calculate URLLC resources= CG + M
    Calculate remaining resources N_f = N - M - CG
    for each eMBB user $e = 1 : E$ **do**
        for each time slot $i = 1 : T$ **do**
            for each resource $j = 1 : N_f$ **do**
                Do this
                Calculate $\mathcal{R}^e_{i,j}$
                Calculate sum data rate of eMBB users using Eq. 5.29
            end for
        end for
    end for
end procedure

5.4 Numerical results and Performance Evaluation

Firstly , the collision probability analysis is performed and then a comparison between Release 16 and optimal resource scheduling schemes is made. The parameters assumed for uplink simulation is shown in Table. 5.1

5.4.1 Collision Probability Analysis

The collision probability in the reserved resources concerning the number of UEs sharing that resource after the first transmission occasion with different resource sizes is shown in Fig. 5.8. It is observed that as the number of users in the network increases, resource sharing increases, which leads to an increase in collision probability with the same number of resource blocks. It is evident

Table 5.1: Parameters

Sub-carrier spacing	60 kHz
repK	1, 2, 4, 8
Collision Probability (P_c)	10^{-3}
Arrival rate (λ)	1.25 * 10^{-4}
BS Power (P_m)[dBm]	20
Carrier BW [MHZ]	40
Path loss exponent	3
RBs per slot (F)	200
Slot length [ms]	0.25
Time domain slots (N)	40
Cell radius [m]	300
eMBB users (E)	10
URLLC users (U)	20, 40, 60, 80, 100

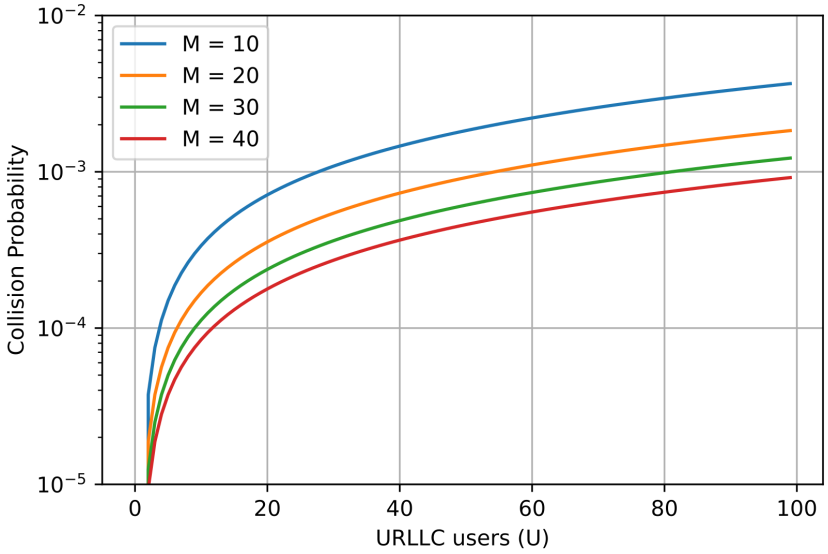

Figure 5.8: Collision probability with respect to N

that an increase in the number of resources results in the successful decoding of data received from users, which increases the reliability of transmissions. Hence, more resources should be assigned to URLLC users in the uplink to achieve URLLC reliability criteria.

The collision probability concerning the size of reserved resources in all transmission occasions for $K = 4$ is shown in Fig. 5.9. From this, to achieve the target collision probability, the number of

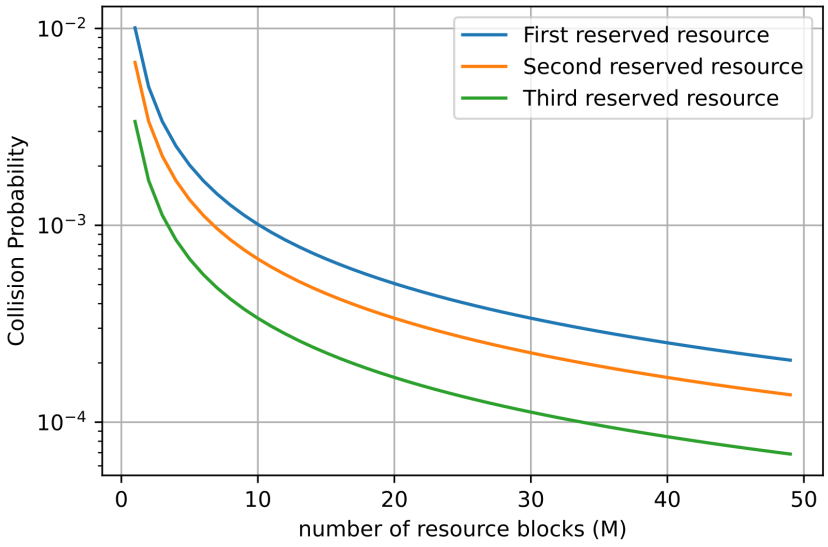

Figure 5.9: Collision probability with respect to number of resource blocks

reserved resources can be calculated when URLLC users in the network are fixed. It is observed that the number of resource blocks decreases based on the position of reserved resource locations in optimal resource allocation for URLLC. As the number of URLLC resources increases to support the same number of users in the network, it is evident that the collision probability decreases.

It was mentioned in the above sections that diversity transmission (multiple repeats of the same packet) improves the reliability of transmissions as the collision probability decreases. Fig. 5.10 is simulated to show the impact of repetitions on reliability. It can be clearly shown that increasing $repK$ suggested by 3GPP, results in a decrease in collision probability. As the number of resources increases, the collision probability also decreases, as was shown earlier.

5.4.2 Comparison of Uplink URLLC Resources Allocation

The number of reserved resources using optimal scheme is calculated based on Eq. 5.27. The number of URLLC users presented in the network is divided into groups as these resources are shared among the UEs. The groups are divided in such a way that these groups of UEs must satisfy the collision probability target (10^{-3}) (from Eq. 5.16) same as reserved resources. In case of multiple configurations by 3GPP, if the gNB configures 4 repetitions, 4 configurations must be set up to guarantee that the UEs may always transmit at the start of a period and reach 4 repetitions as intended. The number of resources required for this scheme is calculated as follows:

- Number of configurations needed for a group of UEs sharing the CG resources are 4.
- Each configuration has 4 CG resources ($repK = 4$) in one period based on SCS. Hence, the number of resources blocks in a period for one group of UEs are 16.

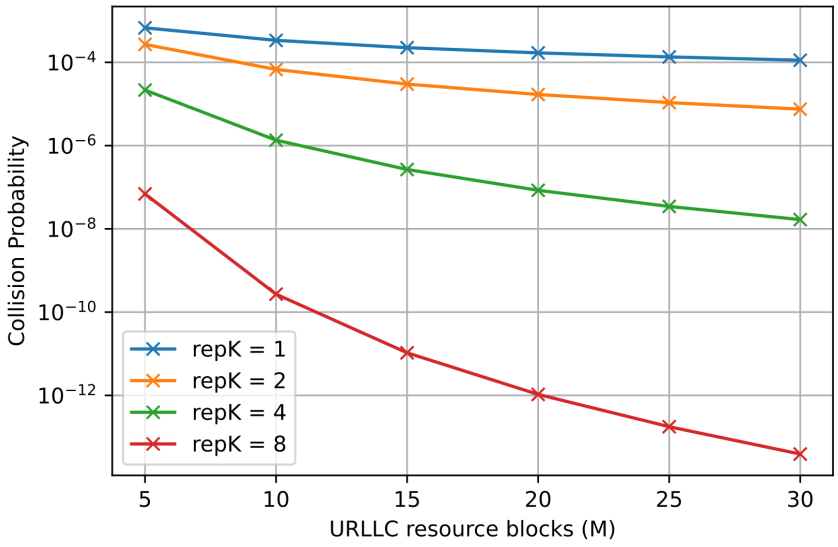

Figure 5.10: Collision probability with respect to URLLC resources (U) varying $repK$

- Based on the number of groups of UEs divided, the total number of resource blocks required to allocate in an interval is calculated.

For example, the sizes of the reserved resources in first, second and third transmission occasions are determined as shown in Table. 5.2 with following parameters: $N = 28$, $repK = 4$, $\lambda = 1.25 * 10^{-4}$. These 28 UEs can be divided into 4 groups of 7 UEs where the collusion probability of the group of UEs is given by Eq. 5.16 as which is approximately equal to 10^{-3}. According to Section. 5.2.1, the number of shared resources in the network is same during repetitions. That implies if the number of reserved resources in first transmission occasion is 10, then the total number of URLLC resources for the same number of users to do repetitions are 40 ($repK = 4$). From this model,the percentage of resources saved as compared to using the same size of resource blocks in for all resources is:

$$\frac{1 - (10 + 7 + 3)}{(10 * 3)} * 100\% = 33.33\%$$

The number of resources regarding allocating multiple configurations (Release 16) results in $4 * 16 = 64$ resource blocks. However the number of CG resources required with optimal scheme are 16 and reserved resources are 20. Hence the resource consumption using the optimal scheme decreases by

$$\frac{36}{64} * 100\% = 56.25\%$$

.

A comparison between these schemes and the variable number of URLLC users in the network is shown in Fig. 5.11. The use of an optimal scheme decreases the number of resources needed for

Table 5.2: Size of reserved resources with repK = 4

Position of reserved resources	1	2	3
Number of resource blocks	10	7	3

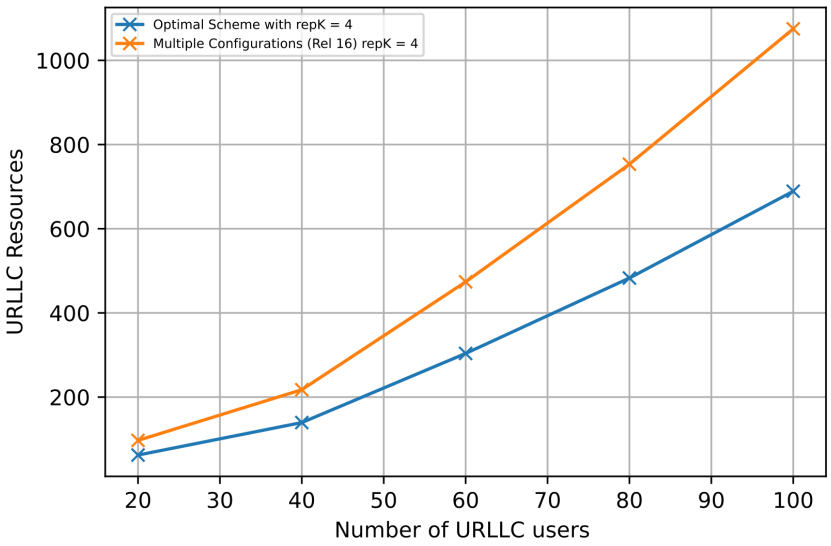

Figure 5.11: No of reserved resource blocks with respect to repetitions

the same number of URLLC users in the network compared to multiple configurations proposed by 3GPP. The conventional scheme in 3GPP Release 15 is that the UE [41] if it cannot do the configured number of repetitions by gNB, must wait until the next period to achieve the URLLC reliability requirement. As a result, in the worst-case scenario with $repK = 4$, 4 transmission occasions occur each period, and the UE must wait for 3 transmission instances, which equates to 3 slots or 0.75 ms with SCS 60kHz. Due to the proximity of this latency to the URLLC requirement of 1 ms, the URLLC UEs are unable to complete the four repeats that were programmed for the following period. Hence the reliability is not ensured therby effecting latency further. The suggested optimal technique, in contrast, enables the UE to begin transmission right away while achieving the configured number of repeats by gNB in the target latency of 1 ms and satisfying the reliability criterion. Hence, the optimal scheme for URLLC resource allocation is considered for the study of eMBB and URLLC multiplexing as least number of resources are reserved for URLLC users.

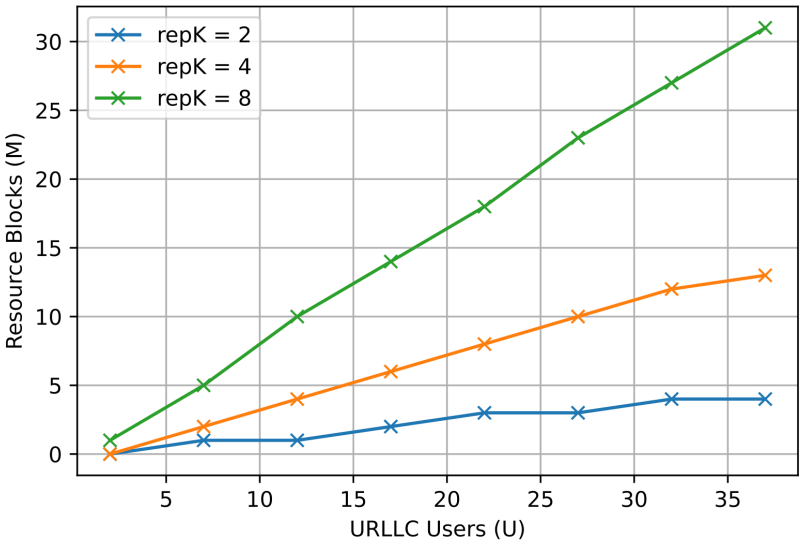

Figure 5.12: Resource Blocks with different URLLC users and $repK$ parameter

Now, for different values of the $repK$ parameter as specified by 3GPP, the number of CG resources and reserved resources (if needed) required to ensure the collision probability is less than 10^{-3} is calculated. If the number of UEs in the system is constant, then the number of resource blocks in each reserved resource location is also calculated. But for the case $K = 1$, there is no requirement of the reserved resources as the optimal scheme is considered for repetitions greater than 1. In-fact, the downlink resource optimization scheduling algorithm can be used for uplink if $repK = 1$. Based on these assumptions, the number of required resource blocks for the collision probability less than 10^{-3} with varying $repK$ and URLLC users is shown in Fig. 5.12. Hence, gNB configures more resources as the number of users in the network increases to maintain target collision probability. It is expected that an increase in the number of repetitions will increase the number of resources used. However, it was shown that the increase in repetitions of packets increases the reliability. When the number of URLLC users in the network increases, the number of resources required increases. Hence, to maintain the target reliability, the gNB has to assign more resources compared to conventional single-shot transmission.

5.4.3 eMBB Data Rate with respect to URLLC Resources

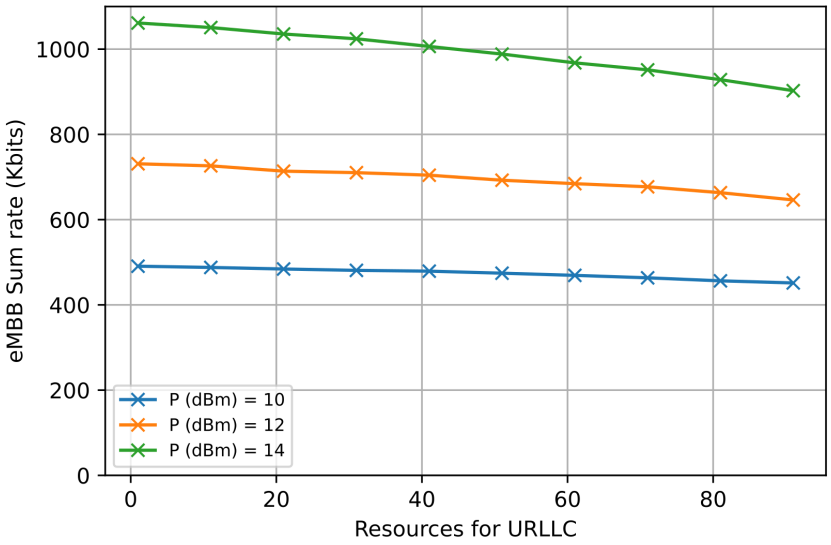

Figure 5.13: Sum rate of eMBB users with varying URLLC Users in the network

Considering optimal resource allocation for URLLC users in the network, the resource allocation for eMBB users is presented in this section. After reserving the resources for URLLC, the remaining resources can be allocated to eMBB users to maximize the data throughput. The sum rate of eMBB users in the network with varying URLLC users and power of base station in the network is shown in Fig. 5.13. It is evident from the plot that as URLLC users in the network increase, the total sum data rate offered by eMBB users decreases. This is because, as URLLC users in the network increases, the number of resources reserved using the optimal scheme increases. This results in a decrease in the eMBB data sum rate. It is also evident that as the power of base station increases, the eMBB data rate increases. This is because. higher power leads to higher reported SNR and UE can choose higher MCS from CQI table which leads to higher data sum rate.

The effect of URLLC users' arrival rate is shown in Fig. 5.14. From the graph, it is clear that the increase in the data arrival rate of URLLC users decreases the data rate of eMBB users. As the arrival rate increase, the number of URLLC resources based on collision probability increases. Then the remaining resources are allocated to eMBB users to maximize the eMBB throughput of users. This causes a decrease in the eMBB rate.

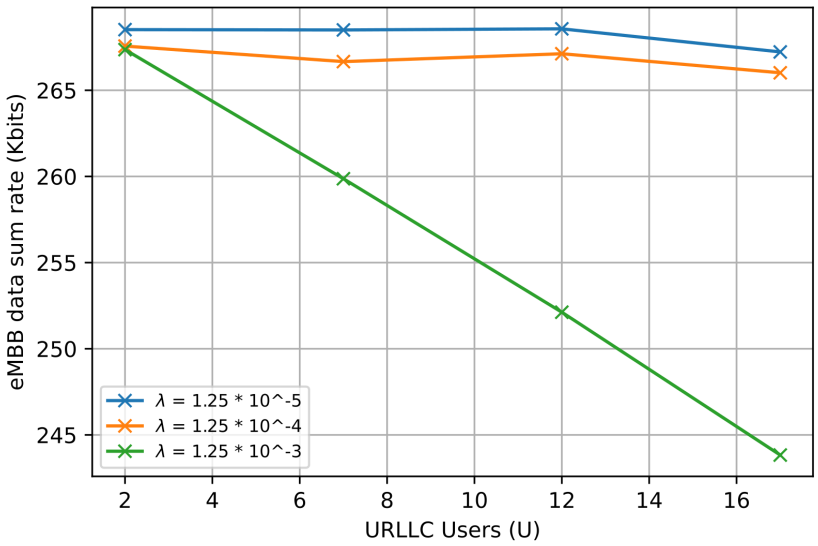

Figure 5.14: Sum rate of eMBB users with varying URLLC Users in the network

Chapter 6

Conclusion and Future Scope

6.1 Conclusion

The URLLC features mentioned in the 5G NR standards have been thoroughly reviewed in this thesis, with an emphasis on the physical layer. New technological concepts are required to meet the demanding specifications, particularly in terms of latency and reliability. Then, we discussed the crucial technological elements or features of the physical layer that would lower latency and boost reliability. Several elements of the physical layer channel and procedure have been newly specified or redesigned to support various use cases for URLLC with a wide variety of latency and reliability requirements: different numerology, K-rep CG transmissions, and network slicing for RAN. By using these enhancements, a resource allocation algorithm is presented in the form of an optimization problem for both uplink and downlink when the network is configured with both eMBB and URLLC users.

Firstly, the dynamic multiplexing of eMBB and URLLC users on the same radio resources using the QoS-aware RAN resource allocation technique. To maximize the network sum rate and meet the diverse requirements of users from two services, the resource allocation problem was described as an AMC-based resource optimization problem. The formulated issue is a very challenging combinatorial mixed-integer non-linear programming optimization issue. The optimization issue was changed into a continuous linear program from previous research and then solved by using the GUROPY tool with Python by relaxing the intractability of AMC and the binary constraint.

Secondly, the presented resource allocation optimization problem is compared with baseline scheduling algorithms. The same AMC-based technique is used in uplink resource allocation. Thirdly, by using uplink CG transmissions, the resource allocation for URLLC users is presented. To meet the requirements for reliability and latency, an optimal resource allocation with reserved resources shared across the UEs enables them to do repeats and arrive at the configured number. Each reserved resource has a size that is optimal for dependability and resource use. A comparison of this optimal scheme and 3GPP solutions is presented. Finally, by using this optimal resource allocation and AMC-based optimized resource allocation discussed for downlink, an uplink resource allocation algorithm is presented in this thesis.

6.2 Future Scope

URLLC's probable implementation strategies from the physical layer's perspective are outlined in this thesis. Both in the theory to comprehend the fundamental limit of URLLC and on relevant models that fit realistic settings, much work is still needed in this area. In addition to the methods we covered, there are several intriguing problems worth investigating, such as the re-configurable URLLC protocol and beam forming tactics for the data. However, the open questions that can be considered as:

- Power allocation to different sub-carriers in both uplink and downlink is assumed to be constant in this thesis. The power is distributed equally to all carriers is considered. However, optimization of power should also should be considered for future work.
- Link adaptation techniques are not discussed in the part of thesis. The CSI from users is assumed to be perfectly decodable at the base station. However, by using outer loop link adaptation (OLLA), algorithms that do not depend on channel estimation can be considered for further work.
- Discussion of the control channel is not considered in this thesis. By adding control channel information to data channels, a new study can be developed using the techniques discussed in this thesis.

List of Figures

List of Tables

Nomenclature

repK	K - repetitions
3GPP	3rd Generation Partnership Project
5G NR	5th Generation New Radio
64-QAM	64 - Quadrature Amplitude Modulation
ACK	Acknowledgement
AMC	Adaptive Modulation Coding
AR	Augmented Reality
ARQ/ HARQ	Automatic Repeat reQuest / Hybrid Automatic Repeat reQuest
AWGN	Additive white Gaussian noise
BLER / BER	Block Error Rate / Bit Error Rate
BS	Base Station
BSR	Buffer Status Report
BW	Bandwidth
CA	Carrier Aggregation
CG	Configured Grant
CN	Core Network
CP	Cyclic Prefix
CQI	Channel Quality Indicator
CSI	Channel State Information
DCI	Downlink control information
DG	Dynamic Grant

DL	Downlink
E2E	End to End
ECDF	Empirical Cumulative Distribution Function
EDS	Equally Distributed Scheduler
eMBB	enhanced Mobile Broad Band
eNodeB	Evolved Node B
FDD	Frequency Division Duplex
GF	Grant Free
gNB	next generation NodeB
HARQ	Hybrid Automatic Repeat reQuest
IoT	Internet of Things
KPI	Key Performance Indicator
LDPC	Low-density parity-check code
LTE	Long Term Evolution
MAC	Medium Access Control
MCS	Modulation and coding scheme
MIMO	Multiple Input Multiple Output
mMTC	massive Machine Type Communication
mmWave	millimeter Wave
NACK	Negative Acknowledgement
OFDM	Orthogonal Frequency Division Multiplexing
PBCH	Physical Broadcase Channel
PDCCH	Physical downlink control channel
PDCP	Packet Data Convergence Protocol
PDF	Probability Density Function
PDSCH	Physical downlink shared channel
PF	Proportional Fair
PHY	Physical Layer

PRACH	Phsycial Random Acess Channel
PRB	physical resource block
PUCCH	Physical uplink control channel
PUSCH	Physical uplink shared channel
QoS	Quality of Service
QPSK	Quadrature phase shift keying
RA	Random Access
RAN	Radio Access Network
RE	Resource Element
RLC	Radio Link Control
RR	Round Robin
RRC	Radio Resource Control
RTT	Round trip time
SCS	Sub-carrier spacing
SDU	Service Data Unit
SINR	Signal-to-interference-plus-noise ratio
SR	Scheduling Request
TB	Transfer Block
TDD	Time Division Duplex
TWG	Tranmission without Grant
UE	User Equipment
UG	Uplink Grant
UL	Uplink
URLLC	Ultra-Reliable and Low-Latency Communication
V2V	Vehicle to Vehicle
V2X	Vehicle to Everything
VR	Virtual Reality

Bibliography

[1] Hyoungju Ji, Sunho Park, Jeongho Yeo, Younsun Kim, Juho Lee, and Byonghyo Shim. Ultra-reliable and low-latency communications in 5g downlink: Physical layer aspects. *IEEE Wireless Communications*, 25(3):124–130, 2018.

[2] Trung-Kien Le, Umer Salim, and Florian Kaltenberger. An overview of physical layer design for ultra-reliable low-latency communications in 3gpp releases 15, 16, and 17. *IEEE access*, 9:433–444, 2020.

[3] Afif Osseiran, Federico Boccardi, Volker Braun, Katsutoshi Kusume, Patrick Marsch, Michal Maternia, Olav Queseth, Malte Schellmann, Hans Schotten, Hidekazu Taoka, et al. Scenarios for 5g mobile and wireless communications: the vision of the metis project. *IEEE communications magazine*, 52(5):26–35, 2014.

[4] Erik Dahlman, Stefan Parkvall, and Johan Skold. *5G NR: The next generation wireless access technology*. Academic Press, 2020.

[5] Hyoungju Ji, Younsun Kim, Juho Lee, Eko Onggosanusi, Younghan Nam, Jianzhong Zhang, Byungju Lee, and Byonghyo Shim. Overview of full-dimension mimo in lte-advanced pro. *IEEE Communications Magazine*, 55(2):176–184, 2016.

[6] Shuangfeng Han, I Chih-Lin, Zhikun Xu, and Corbett Rowell. Large-scale antenna systems with hybrid analog and digital beamforming for millimeter wave 5g. *IEEE Communications Magazine*, 53(1):186–194, 2015.

[7] Carsten Bockelmann, Nuno Pratas, Hosein Nikopour, Kelvin Au, Tommy Svensson, Cedomir Stefanovic, Petar Popovski, and Armin Dekorsy. Massive machine-type communications in 5g: Physical and mac-layer solutions. *IEEE Communications Magazine*, 54(9):59–65, 2016.

[8] Petar Popovski. Ultra-reliable communication in 5g wireless systems. In *1st International Conference on 5G for Ubiquitous Connectivity*, pages 146–151. IEEE, 2014.

[9] 3GPP. Study on new radio access technology Physical layer aspects. Technical Specification (TS) 38.802, 3rd Generation Partnership Project (3GPP), 12 2017. Version 14.2.0.

[10] William Stallings. *5G Wireless: A Comprehensive Introduction*. Addison Wesley, 2021.

[11] 5G. NR; Physical Layer Procedures for Data. document 3GPP TS 38.214, 3GPP, 03 2018. v15.3.0.

[12] Hyunho Lee and Young-Chai Ko. Physical layer enhancements for ultra-reliable low-latency communications in 5g new radio systems. *IEEE Communications Standards Magazine*, 5(4):112–122, 2021.

[13] Ekram Hossain, Mehdi Rasti, and Long Bao Le. *Radio resource management in wireless networks: an engineering approach*. Cambridge University Press, 2017.

[14] 3GPP. Study on new radio access technology Physical layer aspects. Technical Specification (TS) 38.802, 3rd Generation Partnership Project (3GPP), 04 2017. Version 14.1.0.

[15] Frederic Launay. *NG-RAN and 5G-NR: 5G Radio Access Network and Radio Interface*. John Wiley & Sons, 2021.

[16] Bin Liu, Hui Tian, and Lingling Xu. An efficient downlink packet scheduling algorithm for real time traffics in lte systems. In *2013 IEEE 10th Consumer Communications and Networking Conference (CCNC)*, pages 364–369. IEEE, 2013.

[17] Erik Dahlman, Stefan Parkvall, and Johan Skold. *4G: LTE/LTE-advanced for mobile broadband*. Academic press, 2013.

[18] Mohamad Omar Kayali, Zeinab Shmeiss, Haidar Safa, and Wassim El-Hajj. Downlink scheduling in lte: Challenges, improvement, and analysis. In *2017 13th International Wireless Communications and Mobile Computing Conference (IWCMC)*, pages 323–328. IEEE, 2017.

[19] Sameh Musleh, Mahamod Ismail, and Rosdiadee Nordin. Effect of average-throughput window size on proportional fair scheduling for radio resources in lte-a networks. *Journal of Theoretical and Applied Information Technology*, 80(1):179, 2015.

[20] Yan Liu, Yansha Deng, Maged Elkashlan, Arumugam Nallanathan, and George K Karagiannidis. Analyzing grant-free access for urllc service. *IEEE Journal on Selected Areas in Communications*, 39(3):741–755, 2020.

[21] Discussion on Explicit HARQ-ACK Feedback for Configured Grant Transmission. document R1-1903079, , 03 2019. TSG RAN WG1 96.

[22] Adrish Banerjee, DJ Costello, and Thomas E Fuja. Performance of hybrid arq schemes using turbo trellis coded modulation for wireless channels. In *2000 IEEE Wireless Communications and Networking Conference. Conference Record (Cat. No. 00TH8540)*, volume 3, pages 1025–1029. IEEE, 2000.

[23] 3GPP. Radio resource control (rrc) protocol specification. Version 16.4.1.

[24] Jakob Hoydis, Sebastian Cammerer, Fayçal Ait Aoudia, Avinash Vem, Nikolaus Binder, Guillermo Marcus, and Alexander Keller. Sionna: An open-source library for next-generation physical layer research. *arXiv preprint*, Mar. 2022.

[25] Praveen Kumar Korrai, Eva Lagunas, Shree Krishna Sharma, Symeon Chatzinotas, and Björn Ottersten. Slicing based resource allocation for multiplexing of embb and urllc services in 5g wireless networks. In *2019 IEEE 24th International Workshop on Computer Aided Modeling and Design of Communication Links and Networks (CAMAD)*, pages 1–5. IEEE, 2019.

[26] Andrea Goldsmith. *Wireless communications*. Cambridge university press, 2005.

[27] Spyridon Vassilaras, Lazaros Gkatzikis, Nikolaos Liakopoulos, Ioannis N Stiakogiannakis, Meiyu Qi, Lei Shi, Liu Liu, Merouane Debbah, and Georgios S Paschos. The algorithmic aspects of network slicing. *IEEE Communications Magazine*, 55(8):112–119, 2017.

[28] David López-Pérez, Akos Ladányi, Alpár Jüttner, Herve Rivano, and Jie Zhang. Optimization method for the joint allocation of modulation schemes, coding rates, resource blocks and power in self-organizing lte networks. In *2011 Proceedings IEEE INFOCOM*, pages 111–115. IEEE, 2011.

[29] Praveenkumar Korrai, Eva Lagunas, Shree Krishna Sharma, Symeon Chatzinotas, Ashok Bandi, and Björn Ottersten. A ran resource slicing mechanism for multiplexing of embb and urllc services in ofdma based 5g wireless networks. *IEEE Access*, 8:45674–45688, 2020.

[30] Lei You, Mei Song, and Junde Song. Cross-layer optimization for fairness in ofdma cellular networks with fixed relays. In *IEEE GLOBECOM 2008-2008 IEEE Global Telecommunications Conference*, pages 1–6. IEEE, 2008.

[31] Md Shamsul Alam, Jon W Mark, and Xuemin Sherman Shen. Relay selection and resource allocation for multi-user cooperative ofdma networks. *IEEE Transactions on Wireless Communications*, 12(5):2193–2205, 2013.

[32] Wei Yu and Raymond Lui. Dual methods for nonconvex spectrum optimization of multicarrier systems. *IEEE Transactions on communications*, 54(7):1310–1322, 2006.

[33] Gurobi Optimization, LLC. Gurobi Optimizer Reference Manual, 2022.

[34] Mohamed W Nomeir, Yasser Gadallah, and Karim G Seddik. Uplink scheduling for mixed grant-based embb and grant-free urllc traffic in 5g networks. In *2021 17th International Conference on Wireless and Mobile Computing, Networking and Communications (WiMob)*, pages 187–192. IEEE, 2021.

[35] Venkat Yampati. Physical Layer Design for 5G URLLC.

[36] AO Almagrabi, R Ali, D Alghazzawi, A AlBarakati, and T Khurshaid. A poisson process-based random access channel for 5g and beyond networks. mathematics 2021, 9, 508, 2021.

[37] Semi-Persistent Scheduling for 5G New Radio URLLC. document R1 -167309, 3GPP TSG-RAN WG1 86, 08 2016.

[38] K.Daniel Wong. *Fundamentals of Wireless Communication Engineering Technologies*. Wiley, 2012.

[39] Bikramjit Singh, Olav Tirkkonen, Zexian Li, and Mikko A Uusitalo. Contention-based access for ultra-reliable low latency uplink transmissions. *IEEE Wireless Communications Letters*, 7(2):182–185, 2017.

[40] UL Grant-Free Transmission for URLLC. document R1-1705246, , 04 2017.

[41] Trung-Kien Le, Umer Salim, and Florian Kaltenberger. Optimal reserved resources to ensure the repetitions in ultra-reliable low-latency communication uplink grant-free transmission. In *2019 European Conference on Networks and Communications (EuCNC)*, pages 554–558. IEEE, 2019.

[42] 3GPP R1-1812162. Enhancement of configured grant for nr urllc. 2018.

[43] 3GPP R1-1812226. Enhanced ul configured grant transmissions. 2018.

[44] Edwin KP Chong and Stanislaw H Zak. *An introduction to optimization*, volume 75. John Wiley & Sons, 2013.

[45] Eric W Weisstein. Binomial theorem. *https://mathworld. wolfram. com/*, 2002.

Appendix A

A.1 OFDM

The output of multi-carrier transmitter in OFDM is given as [13]

$$y(t) = \sum_{v=0}^{\mathcal{N}-1} Y_v \, e^{j2\pi f_v t} \tag{A.1}$$

where Y_v is a complex variable with constellation (based on modulation) with data length $\mathcal{N}$ and f_v is the vth carrier frequency for Y_v. Digital transmitters and receivers are used in modern communication systems most of the time. The output of a digital transmitter will be produced using sampling and quantization converting from analog. The transmitted signal is produced by the transmitter by combining the modulator outputs. The output of digital transmitter with multi-carrier modulation is

$$y(mT_s) = \sum_{v=0}^{\mathcal{N}-1} Y_v e^{j2\pi f_v m T_s} \tag{A.2}$$

where T_s is the sampling interval. Additionally, if the carrier frequencies have a constant frequency spacing of f_s i.e $f_v = vf_s = \frac{vf_s}{\mathcal{N}T_s}$, where v = 0,1,2,.$\mathcal{N}-1$ and are evenly spaced apart in the frequency domain, then the output of the transmitter is given in Eq. A.3. The constant spacing between the carriers is known as sub-carrier spacing.

$$y_m = y(mT_s) = \sum_{v=0}^{\mathcal{N}-1} Y_m e^{\frac{j2\pi vm}{\mathcal{N}}} \tag{A.3}$$

Appendix B

B.1 Optimization

The study of optimization, a vast and expanding area of mathematics, focuses on issues related to decision-making. The optimization techniques ultimately aim to either minimize the amount of effort or maximize the desired benefit. The following is a formula for an optimization or mathematical programming problem [44]:

$$\begin{aligned} \min_{x} \quad & f(x) \\ \text{subject to} \quad & d_j(x) \leq 0, j = 1, 2, ..., m \\ & g_j(x) = 0, j = 1, 2, .., p; \end{aligned} \tag{B.1}$$

where x is an n-dimensional design vector, $f(x)$ is the objective or merit function, and $d_j(x)$, $m_j(x)$ are the constraints which are inequality and equality respectively. The above optimization problem is known as constrained optimization as it is subjected to different constraints and it can be unconstrained problem without any constraints. Following the modeling phase, one can solve an optimization problem using a variety of techniques. There isn't a single way to tackle optimization problems; rather, there are a variety of algorithms for various problems like linear optimization, convex optimization etc. In order to solve an optimization problem, the steps need to be considered are [44]:

- utilizing optimization methods to create optimization models
- Identify whether or not the problem is solvable.
- transform optimization issues into tractable forms
- selecting the appropriate algorithms to solve the models

Based on the types of equations used for the constraints and the objective function, optimization problems can be categorized as linear, nonlinear, geometric, and quadratic programming problems. Since there are numerous specialized methods available for the effective solution of a certain class of problems, this classification is quite helpful from a computational perspective [44].

B.2 Convex Optimization

A function's convexity is crucial for understanding an optimization problem. If the line segment connecting any two points in a set C also lies in C, then the set C is convex. This indicates that for any two points x_1,x_2 of C and $\theta \in R$, such that $0 \leq \theta \leq 1$,

$$\theta x_1 + (1-\theta)x_2 \in C \tag{B.2}$$

In the case of any pair of points, a function $F(X)$ is said to be convex for all λ, $0 \leq \lambda \leq 1$,

$$F[\lambda X2 + (1-\lambda)X1] \leq \lambda F(X2) + (1-\lambda)F(X1) \tag{B.3}$$

A convex function's negative is a concave function, and vice versa. In fact, if the problem is shown to be convex, it will become considerably easier. The fact that any local minimum is also a global minimum is the key characteristic that draws attention to convex problems [13]. Even when a problem isn't convex, it can be converted to convex by some relaxation techniques to get a lower convex bound on the original issue. Different algorithms like Newton's method and gradient descent methods are utilized to find the optimal point of an optimization problem. The problem from Eq. B.1 is convex optimization problem when the objective function is differentiable and the constraint functions are convex which is also refereed as **primal problem**. Utilizing the Lagrangian dual function, which is always a convex function, is one technique to get a lower bound of a problem. Considering the Eq. B.1, a Lagrangian function is defined as follows:

$$L(x,\lambda,\mu) = f(x) + \sum_{j=1}^{m} \lambda_i d_i(x) + \sum_{j=1}^{p} \mu_i h_i(x) \tag{B.4}$$

where $\lambda_i \geq 0$ and β_i are the Lagrangian multipliers for inequality and equality constraints respectively for the optimization problem. The Lagrangian dual function is defined as follows:

$$g(\lambda,\mu) = \max_{x} \quad L(x,\lambda,\mu) \tag{B.5}$$

From [13], the proposition defines if p is the optimal value of primal problem , then for any $\lambda \geq 0$ and any μ results in g(λ,μ) $\leq p$. A Lagrangian dual problem is defined as:

$$\begin{aligned} \max_{\lambda} \quad & g(\lambda,\mu) \\ \text{subject to} \quad & \lambda > 0 \end{aligned} \tag{B.6}$$

If d is the optimal solution for the Lagranian dual optimization problem, then the difference $p-d$ is knowns as optimal duality gap. The dual optimization problem is easy to solve than primal one[13]. The strong duality holds when the difference is zero. Hence the resource allocation issue which is converted to an optimization problem can be solved using above process by converting it to convex optimization problem.

B.3 Rayleigh Fading Channel

The constant need to assess the potential for their performance enhancement is brought on by the ongoing development of various wireless communication systems. Unfortunately, there are a

number of side effects and disadvantages associated with signal propagation in the wireless medium, including multipath fading and noise. Assuming two Gaussian distributed random variables with zero means, X_1 and X_2, each having a variance of

$$\epsilon^2$$

, then the Eq. B.7 represents Rayleigh distributed random variable.

$$r = \sqrt{X_1^2 + X_2^2} \tag{B.7}$$

The random variable with probability density function (pdf) is given as follows:

$$f_R(r) = \frac{2r}{\kappa} exp(-\frac{r^2}{\kappa}) \tag{B.8}$$

where $\kappa = 2\epsilon^2$ represents the average signal power which can be given by expectation of the random statical process ($E(r^2)$).

B.4 Binominal Theorem

For real or complex a, b, and non-negative integer q, the Binomial Theorem states that [45]

$$(a+b)^q = \sum_{r=0}^{q} \binom{q}{r} a^{q-r} b^r \tag{B.9}$$

where $\binom{q}{r} = \frac{q!}{r!(q-r)!}$ represented as a binominal coefficient.