Impact of slicing on radio resource management in 5G for vehicular URLLC and eMBB

Nathalie Naddeh

INSTITUT POLYTECHNIQUE DE PARIS

NNT : 2022IPPAS021

Thèse de doctorat

Impact of slicing on radio resource management in 5G for vehicular URLLC and eMBB

Thése de doctorat de l'Institut Polytechnique de Paris
préparée à Télécom SudParis

École doctorale n°626 École doctorale de l'Institut Polytechnique de Paris (EDIPP)
Spécialité de doctorat : Réseaux, informations et communications

Thèse présentée et soutenue à Palaiseau, le 13/12/2022, par

NATHALIE NADDEH

Composition du Jury :

Kinda Khawam Maître de Conférences/HDR, Université de Versailles St-Quentin en Yvelines (DAVID)	Rapporteur
Thi-Mai-Trang Nguyen Professeur des universités/HDR, Univesrité Sorbonne Paris Nord (L2TI)	Rapporteur
Loutfi Nuaymi Professeur, IMT Atlantique (SRCD)	Président et Examinateur
Walid Ben-Ameur Professeur, Telecom SudParis (SAMOVAR)	Examinateur
Tijani Chahed Professeur, Télécom SudParis (SAMOVAR)	Directeur de thèse
Sana Ben Jemaa Ingénieur Chercheur, Orange Labs (REP)	Co-encadrante
Salah Eddine El Ayoubi Professeur, Centrale Supelec (L2S)	Invité

Résumé

La 5G-NR (Fifth Generation-New Radio) a introduit le concept de slicing pour cibler différents types de services. Nous considérons dans cette thèse le trafic véhiculaire, les véhicules envoyant deux types de flux : eMBB (enhanced Mobile BroadBand) et URLLC (Ultra-Reliable and Low Latency Communications). Ces flux sont acheminés en deux slices différents, la première cherchant à garantir et/ou maximiser le débit, tandis que la seconde doit répondre à de fortes contraintes de QoS(Quality of Service) en termes de délai, de l'ordre de 1ms, et de fiabilité, sur de l'ordre de 99.999%. Ces slices avec des profils de trafic et des exigences de QoS hétérogènes doivent partager la même infrastructure physique.

Cette thèse vise à proposer de nouveaux schémas d'allocation de ressources pour satisfaire les exigences strictes de qualité de service de l'URLLC sans impacter trop le trafic eMBB. L'un des principaux défis est le moment où les ressources initialement réservées à l'eMBB doivent être allouées à l'arrivée de nouveaux flux URLLC. En raison de l'utilisation de différentes numérologies, ces ressources doivent être reconfigurées, ce qui ajoute un délai supplémentaire de l'ordre de 80 ms, ce qui dépasse le budget de délai URLLC. Pour répondre à ce problème de délai, nous proposons des schémas proactifs de réservation de ressources pour URLLC qui anticipent l'arrivée des véhicules dans une cellule et (re-)configurent le slice avant leur arrivée effective dans la cellule. Ces approches permettent de répondre aux exigences de délai et de débit du trafic URLLC et eMBB des véhicules, respectivement.

Nous introduisons en outre un modèle de dimensionnement inter-slice qui prend en compte les conditions radio et les trajectoires de l'utilisateur dans le réseau, ce qui permet de prendre en compte les MCS (Modulation and Coding Scheme) des utilisateurs. Ce faisant, nous obtenons une meilleure allocation des ressources grâce à une optimisation plus fine. Nos résultats montrent que nous sommes en mesure de satisfaire les exigences de trafic avec une meilleure utilisation des ressources. Finalement, nous étudions

un modèle de dimensionnement alternatif basé sur des bornes de grande déviation. Nous analysons la queue du système correspondant à la région de perte URLLC. Nous considérons deux approches : avec et sans mise en file d'attente de paquets. Nous observons que les grandes limites d'écart entraînent une surréservation légèrement supérieure à l'approche susmentionnée lorsqu'elle est appliquée à l'URLLC, avec l'avantage du calcul instantané des ressources nécessaires.

Abstract

The Fifth Generation-New Radio (5G-NR) introduced the concept of slicing to target different types of services. We consider in this thesis vehicular traffic, with vehicles sending two types of flows: enhanced Mobile BroadBand (eMBB) and Ultra-Reliable and Low Latency Communications (URLLC). These flows are transported in two different slices, the former trying to guarantee and/or maximize the throughput, while the latter has to meet stringent Quality of Service (QoS) constraints in terms of delay, on the order of 1ms, and reliability, on the order of 99,999%. These slices with heterogeneous traffic profiles and QoS requirements must share the same physical infrastructure.

This thesis aims to propose new resource allocation schemes to satisfy URLLC stringent QoS requirements without impacting too much eMBB traffic. One main challenge is when resources initially reserved for eMBB must be allocated to the arrival of new URLLC flow. Due to using different numerologies, these resources need to be reconfigured, adding extra delay on the order of 80ms, which exceeds the URLLC delay budget. To respond to this delay problem, we propose proactive resource reservation schemes for URLLC which anticipates the vehicles' arrival in a cell and (re-)configures the slice before their effective arrival in the cell. These approaches enable to meet the delay and throughput requirements of vehicular URLLC and eMBB traffic, respectively.

We additionally introduce an inter-slice dimensioning model that considers user's radio conditions and trajectories in the network, which enables taking into consideration users Modulation and Coding Schemes (MCS). By doing so, we achieve a better resource allocation through finer optimization. Our results show that we are able to satisfy traffic requirements with a better resource utilization. Eventually, we investigate an alternative dimensioning model based on large deviation bounds. We analyze the tail of the system corresponding to the URLLC outage region. We consider two approaches:

with and without packet queuing. We observe that large deviation bounds result in slightly more over-reservation than the aforementioned approach when applied to URLLC, with the advantage of instantaneous computation of the needed resources.

Acknowledgments

First and foremost, I would like to thank the members of the jury for accepting to be a part of this thesis. Second, my appreciation goes to my supervisors, Prof. Tijani Chahed, Prof. Salah Eddine El Ayoubi, and Dr. Sana Ben Jemaa. Working with them for the last three years was an honor and an enlightening experience. Even though a global pandemic has stalled our work rhythm, they remained very supportive and adapted our work to this situation. Their dedication, commitment, and love for their work have always inspired me. Their valuable advice and mentoring have set me on the right track to present this work and prepared me for the next chapter. So I would like to hereby state my gratefulness for the attention they gave.

I also would like to thank Vincent Diascorn, my manager at Orange Labs, for his constant attention and interest in this thesis, and his effort for my team integration, as well as my teammates. My integration into the team was easy because I had the luck to meet such welcoming and easy-going people. The environment at Orange Labs has contributed immensely to appreciating my work. I would also like to thank my friends at Orange Labs with whom I enjoyed coffee breaks and discussing life: Mira, Imene, Romain, Youssef, Ali, Antoine, Mohamad, Rita, Amel, and Marie. You made my life so much easier and enjoyable during these years. I will always appreciate our time spent together and all the laughs.

I would like to express, on a personal note, my gratitude to Mariah El Asmar and Cléa el Murr for their friendship. Even with the distance, you were always there for me and helped me through these years. My life is a lot better with you in it. I must thank my parents for supporting me throughout the years and for being my most ardent fan. Words fail to express how much I love you, thank you for making me who I am today.

Last but most importantly, I would also like to extend my deepest gratitude and appreciation to my husband, Henry Jello, who has always been there for me. He is always pushing me to be my better self and do my best. I am so

proud to be your wife and I love you.

Contents

List of Figures

List of Tables

List of Algorithms

Acronyms

3GPP	Third Generation Partnership Project.
4G	Fourth Generation.
5G-EIR	5G-Equipment Identity Register.
5G-NR	Fifth Generation-New Radio.
5GC	5G Core.
AF	Application Function.
AI	Artificial Intelligence.
AMF	Access and Mobility Management Function.
API	Application Programming Interfaces.
AR	Augmented Reality.
AUSF	Authentication Server Function.
BLER	Block Error Rate.
BSRA	Bandwidth Slicing and Resource Allocation.
BSS	Business Support Systems.
BWP	Bandwidth Part.
CP	Cyclic Prefix.
CP	Control Plane.
CQI	Channel Quality Indicator.
CSMF	Communication Service Management Function.
DCI	Data Center Interconnect.
DL	Down-link.
DN	Data Network.
DRL	Deep Reinforcement Learning.

E-UTRAN	Evolved-UMTS Terrestrial Radio Access Network.
E2E	End-to-End.
eMBB	enhanced Mobile BroadBand.
FCFS	First-Come-First-Served.
FTP	File Transfer Protocol.
gNodeB	Next Generation Node B.
HARQ	Hybrid Automatic Repeat reQuest.
HSS	Home Subscriber Server.
IUT-R	International Telecommunication Union Radiocommunication.
KPI	Key Performance Indicator.
LPWA	Low Power Wide Area.
LTE	Long Term Evolution.
MCS	Modulation and Coding Scheme.
MDP	Markov Decision Process.
MISO	Multiple-Input-Single-Output.
ML	Machine Learning.
MME	Mobility Management Entity.
mMTC	massive Machine Type Communications.
NB-IoT	NarrowBand-Internet of Things.
NEF	Network Exposure Function.
NFV	Network Function Virtualization.
NG	Network Getaway.
NG-C	Network Getaway User.
ng-eNB	next generation enhanced Node B.
NG-U	Network Getaway Control.
NR	New Radio.
NRF	NF Repository Function.
NS 5G-NR	Non-Standalone Fifth Generation-New Radio.
NSI	Network Slice Instance.
NSMF	Network Slice Management Function.

NSSAI	Network Slice Selection Assistance Information.
NSSF	Network Slice Selection Function.
NSSMF	Network Slice Subnet Management Function.
OFDM	Orthogonal Frequency Division Multiplex.
PCF	Policy Control Function.
PCRF	Policy and Charging Rule Function.
PDN	Packet Data Network.
PLMN	Public Land Mobile Network.
Q-Learning	Quality-Learning.
QoS	Quality of Service.
RAC	Radio Admission Control.
RAN	Radio Access Network.
RAT	Radio Access Technologies.
RB	Resource Block.
RF	Radio Frame.
RL	Reinforcement Learning.
RRC	Radio Resource Control.
RRM	Radio Resource Management.
S-NSSAI	Single-Network Slice Selection Assistance Information.
SBI	Service Based Interfaces.
SCS	Sub-Carrier Spacing.
SD	Slice Differentiator.
SDN	Software-Defined Networking.
SEPP	Security Edge Protection Proxy.
SGSN	Serving GPRS Support Node.
SGW	Serving Gateway.
SINR	Signal to Interference Noise Ratio.
SLA	Service Level Agreement.
SMF	Session Management Function.
SNR	Signal to Noise Ratio.
SPS	Semi-Persistent Scheduling.
SST	Slice/Service Type.
sTTI	short TTI.

TTI	Transmission Time Interval.
UDM	Unified Data Management.
UDR	Unified Data Repository.
UDSF	Unstructured Data Storage Function.
UE	User Equipment.
UL	Up-Link.
UP	User Plane.
UPF	User Plane Function.
URLLC	Ultra-Reliable and Low Latency Communications.
V2X	Vehicle-to-everything.
VR	Virtual Reality.

Résumé Générale

Introduction

La quatrième génération (4G) et le Long Term Evolution (LTE) ont été une nouveauté en termes de débits de données plus élevés et l'adoption de la technique de commutation par paquets uniquement ainsi que le changement global du réseau central. La croissance du trafic et l'augmentation des exigences de QoS ont rendu la 4G insuffisante en termes de débits de données, de latence et de flexibilité. Il était donc nécessaire de trouver une solution adaptée à cette diversité. Les réseaux 5G ont été introduit dans le but de répondre à cette croissance et diversité d'utilisateurs. La 5G offre des débits de données plus élevés, des latences plus faibles, une plus grande flexibilité, et évolutivité, selon le besoin des différents services. La 3GPP a définit 3 cas d'usages qui regroupent les differents services.

D'une part, la 5G offre des services à haut débit mobile enhanced Mobile BroadBand (eMBB), nécessitant une couverture radio sans faille et des débits de données élevés. Parmi les cas d'utilisation envisagés, nous mentionnons la réalité augmentée, l'expérience d'événements immersifs et les vidéos 8K. D'autre part, les systèmes 5G vont également initier une évolution disruptive, définissant de nouveaux cas d'usages. Ces services sont divisés en 2 catégories. Tout d'abord, la catégorie des massive Machine Type Communications (mMTC), qui demande une augmentation exponentielle du nombre de dispositifs connectés, ayant une faible éxigence en terme de débit et de latence. Deuxièmement, la famille de service Ultra-Reliable and Low Latency Communications (URLLC) définit des exigences strictes en termes de latence et de fiabilité pour des applications dans le secteur médical, industrie 4.0 et des voitures autonomes par exemple.

Pour provisioner ces nouveaux services sous la même infrastructure physique, un slicing de réseau de bout en bout, a été introduit. Le slicing du réseau est un nouveau concept introduit pour permettre aux opérateurs de cibler

de nouveaux marchés appelés verticaux qui bénéficieront des trois classes de services. Une slice est une collection de ressources réseau, sélectionnées pour satisfaire les demandes, en termes de Quality of Service (QoS).

Le slicing du réseau permet à l'opérateur de créer des réseaux personnalisés et de fournir des solutions optimisées pour différents scénarios. Le slicing vise à introduire de la flexibilité et une meilleure utilisation des ressources réseau. Il offre les ressources réseau nécessaires pour répondre aux exigences des slices actives. La 3GPP l'a défini comme une solution permettant d'héberger différents types de services sur la même infrastructure physique, chaque slice étant considérée comme une infrastructure virtuelle.

La coexistence des différents slices sous une infrastructure unique est considérée comme un processus très complex, surtout sur la partie Radio Access Network (RAN) du réseau. Leur multiplexage sur une même infrastructure doit se faire de manière flexible, efficace et optimale. La gestion de l'allocation des ressources radio est certainement un problème de recherche crucial, encore ouvert. Avec l'émergence de plusieurs technologies qui s'ajoutent au RAN, les procédures Radio Resource Management (RRM) deviennent plus sophistiquées et complexes.

Motivation

Dans le contexte du scénario véhiculaire, l'URLLC est véhiculée par les technologies Vehicle-to-everything (V2X). Les technologies V2X ont été introduites dans le 5G-NR pour couvrir diverses applications, telles que la conduite autonome, le peloton de véhicules, les applications embarquées critiques comme les ambulances connectées, etc [1]. Nous considérons dans cette thèse le trafic véhiculaire, les véhicules envoyant deux types de flux : eMBB et URLLC. Ces flux sont acheminés en deux slices différents, la première cherchant à garantir et/ou maximiser le débit, tandis que la seconde doit répondre à de fortes contraintes de QoS en termes de délai, de l'ordre de 1ms, et de fiabilité, sur de l'ordre de 99.999%.

Cette coexistence nécessite un management d'allocation de ressource, qui peut être représenter sur deux axes: Allocation de ressources Intra et Inter slice. Au niveau d'allocation intra-slice, l'ordonnanceur alloue les ressources disponibles entre les utilisateurs attachés à un slice spécifique. L'ordonnanceur intra-slice utilise des schémas d'ordonnancement classiques comme Round robin et proportionnal fair. En ce qui concerne la gestion intra-ressource pour le service URLLC, pour atteindre les faibles latences, le besoin de numérologie

spécifique a été adopté par la 3GPP, avec un short TTI (sTTI) et un codage de canal robuste. Cette numérologie doit être sélectionnée sur la base d'un algorithme qui prend en entrée la moyenne Signal to Noise Ratio (SNR), l'effet Doppler et l'étalement des retards [2].

L'allocation de ressources inter-slice résout le problème de la planification des ressources radio dans le RAN en fonction d'un SLA spécifique de chaque slice. L'ordonnancement des ressources entre les slices est considéré l'un des principaux défis. Il doit être intelligent, flexible et cognitive pour répondre aux diverses exigences de qualité de service des flux 5G.

L'un des plus gros défit est le moment où les ressources initialement réservées à l'eMBB doivent être allouées à l'arrivée de nouveaux flux URLLC, si les ressources ne sont pas suffisantes. Dans ce cas, une nouvelle composante de délai s'ajoute à la latence radio, qui est la latence de file d'attente. En raison de l'utilisation de différentes numérologies, ces ressources doivent être reconfigurées en s'adaptant à cette numérologie, ce qui ajoute un délai de file d'attente supplémentaire de l'ordre de 80 ms. Ce délai dépasse le budget de latence URLLC. Pour répondre à ce problème de délai, nous proposons des schémas proactifs de réservation de ressources pour URLLC qui anticipent l'arrivée des véhicules dans une cellule et (re-)configurent le slice avant leur arrivée effective dans la cellule. Ces approches permettent de répondre aux exigences de délai et de débit du trafic URLLC et eMBB des véhicules, respectivement.

Dans l'objectif est de mettre en place des mécanismes de gestion des ressources radio nécessaires pour contrôler ces différentes slices et appliquer les schémas proactifs proposés, nous proposons deux modèles dimensionnants, exploitant les conditions radio des utilisateurs obtenues par le réseau. Un premier basé sur la mise en file d'attente des paquets et un second basé sur les grandes limites de déviation. Nous allouons dynamiquement les ressources entre les slices avec une méthode de planification inter-slices qui garantit le SLA de l'URLLC véhiculaire tout en limitant la dégradation des performances eMBB. Cette méthode de planification prend en compte l'allocation proactive pour anticiper le dimensionnement. Les cas d'usage sont simulés sur un simulateur de système 5G décrit dans 3.2.2.

Structure de la thèse

Durant cette thèse, notre but était de proposer de nouveaux schémas d'allocation de ressources pour satisfaire les exigences strictes de qualité de service de

l'URLLC sans impacter trop le trafic eMBB. Pour répondre à cette problématique, nous avons structuré notre travail comme suit :

- Le chapitre 2 définit certains des outils de slicing 5G, notamment la création de slices, la gestion du slicing et les procédures RRM pour l'allocation des ressources. Nous introduisons également le problème du délai de reconfiguration du slice dans le cas de la coexistence de plusieurs slices, ce qui motive nos approches proactives de réservation de ressources.
- Dans le chapitre 3, nous énumérons certains des travaux qui traitent du sujet du multiplexage de slices et des procédures RRM en cas de slicing. Nous décrivons le modèle de système adopté dans la thèse ainsi que le simulateur développé. Nous détaillons ensuite les approches proactives de réservation de ressources que nous proposons pour résoudre le problème de retard supplémentaire dû à la reconfiguration.
- Au chapitre 4, nous décrivons le modèle de dimensionnement que nous proposons pour obtenir une allocation optimale des ressources entre slices. Ce modèle est basé sur les statistiques radio et le SLA du slice URLLC. Nous appliquons ce modèle sur un simulateur de réseau 5G et analysons le gain de performance pour les slices URLLC et l'impact sur le débit des utilisateurs eMBB.
- Dans le chapitre 5, pour résoudre le problème des simulations chronophages, nous introduisons un modèle de dimensionnement basé sur des bornes de grande déviation qui donnent des équations closed form. Nous comparons le modèle analytique avec des simulations numériques et des travaux éxistants, en utilisant notre simulateur de système 5G.
- Le chapitre 6 conclut nos travaux de thèse et donnent quelques indications sur les perspectives de travaux futures, à court et à long terme.

Chapter 1

Introduction

In this introduction, we give a global view of the Fifth Generation-New Radio (5G-NR) and the concept of slicing as well as new verticals. The description of the new use-cases is detailed, with a more in depth description of the URLLC use case which is the main focus of our work. The co-existence of this type of service with others is the most challenging part since the RAN and the core network need to adapt their procedures and functionalities to respond to the services' demands. The operator and the verticals should reach an agreement called the Service Level Agreement (SLA) that aims to satisfy the vertical's needs while taking into account the operator's capacity. The management of the resources between these slices is the main focus of our work, as we shall detail in the contributions listing.

1.1 Context

1.1.1 5G-NR and network slicing

5G-NR has brought new technologies, new modes of connectivity, and new ways to configure and optimize the network [3]. Its main purpose is to address the new demands of the introduced services and enhance the performance of the existing ones. It is expected to be able to provide optimized and flexible support for a variety of different communication services, different traffic loads, and different end-user communities.

Therefore, in 5G, there is a need to push the envelope of performance to provide, where needed, much greater throughput, much lower latency, ultra-high reliability, much higher connectivity density and higher mobility

range. This enhanced performance is expected to be provided along with the capability to control a highly heterogeneous environment, and the capability to, among others, ensure security, trust and privacy.

5G-NR is being designed to support different types of services under 3 main use cases (Figure 1.1):

- eMBB provides higher bandwidth and data rates (up to 1 Gbps) and better latency for newer applications such as 4K media, Augmented Reality (AR) and Virtual Reality (VR). We consider the eMBB traffic as the extension of the 4G broadband service. Its characteristics are defined with its large payload and long activation time.
- URLLC is required to support applications with very high reliability (99.9999%) and low latency (1 ms) [4]. The transmissions of URLLC packets are intermittent and short, with a small payload. Their traffic can be sporadic or periodic.
- mMTC, which has been already developed as part of Third Generation Partnership Project (3GPP) Release 13/14 Low Power Wide Area (LPWA) technologies including NarrowBand-Internet of Things (NB-IoT), requires higher connectivity of devices for smart cities and other IoT applications (up to 1 Million connections/km^2) [5]. The randomness of activity of this type of service makes the application of a priori resource allocation not feasible.

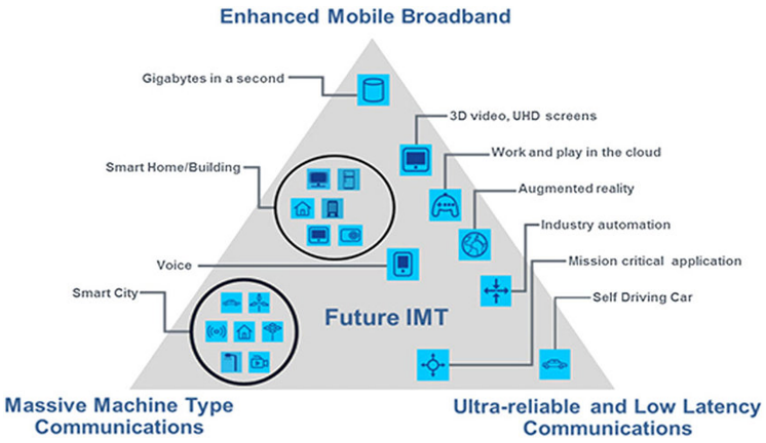

Figure 1.1: 5G use cases(source: IUT-R, 2015)

To support these services under the same physical infrastructure, an end-to-end, so-called network slicing was introduced. Network slicing is a new

concept introduced to allow operators to target new markets called verticals that will profit from the three above-mentioned service classes: eMBB, URLLC and mMTC. A slice is a collection of network resources, selected to satisfy the demands, in terms of QoS, of the service(s) to be delivered by the slice [6]. We create several slices depending on their requirements and attach them to a use case (from the three above-mentioned ones).

Network slicing enables the operator to create customized networks and to provide optimized solutions for different market scenarios. Slicing aims to introduce flexibility and better network resources utilization. It offers the network resources necessary to fulfill the requirements of active slices. 3GPP defined it as a solution enabling the accommodation of different types of services on the same physical infrastructure, while each slice is considered as one virtual infrastructure.

This flexibility however comes at the expense of added network management complexity. Automation of network slice management processes is, therefore, crucial to effectively use the excess of features provided by 5G and the flexibility brought by network slicing.

1.1.2 Service Level Agreement

The novelty of slicing is not reduced to differentiating groups of users. The absolute novelty comes from this commitment to the QoS. It guarantees each slice by a contract that binds the operator to the vertical who owns the content of the slice. 3GPP defines this contract as the SLA. The business level defines the SLA contract. One key differentiation of network slicing is that they can support multiple services with their individual SLAs.

A "good" definition of SLA is essential to allow the operator to manage its network and meet its commitments properly. The SLA could, for example, be defined in terms of the coverage area, data throughput, required latency, etc., to name just a few. But it could also be more restrictive, requiring a guaranteed allocation of network resources like spectrum band. For these scarce resources, it is not feasible to have dedicated allocations for a large number of network slices. So while defining these SLAs, we need to consider the requirements of all the operational network slices and the future potential extensions. The more the SLA description is accurate, the more the slice management can be efficient. If the slice deployment is local, we may need to have a dedicated infrastructure with a dedicated spectrum. Hence the SLA will concern a commitment to the resources, especially if the tenant manages the local network. If the operator is responsible for network management,

then the SLA can be expressed in terms of QoS guarantees. The operator can commit to guaranteed quality in a given coverage area and given traffic demand.

Its variation at the level of the radio segment will allow this segment to set its objectives and implement the mechanisms necessary to achieve them. Knowing how to measure the achieved objectives at the radio level is important. Differentiated monitoring by the slice is then essential.

This service differentiation would require different RAN architectures and designs to support these verticals, which is very challenging economically and environmentally. The new 5G RAN is the result of these challenges. It introduces new functionalities that keep up with the diversity of the customers. For example, new scheduling techniques are proposed to support slicing, such as preemptive and semi-persistent scheduling.

1.1.3 URLLC slicing concept

The 3GPP Release 15 introduced the URLLC service to address the requirements of ITU-R M.2083. It is a primary enabler for several unique use cases in manufacturing, energy transmission, transportation and healthcare. With the need to support End-to-End (E2E) latency as low as 5ms, the delay budget for individual interfaces can be as low as 1ms. We must consider optimization at every step of the uplink and downlink.

In the context of the vehicular scenario, URLLC is conveyed through V2X technologies. V2X technologies have been introduced in the 5G-NR to cover various applications, such as autonomous driving, vehicle platooning, mission-critical onboard applications like connected ambulances, etc [1]. Their coexistence with other services, especially eMBB using the same infrastructure, is managed thanks to the network slicing paradigm. Their QoS or SLA requirements in terms of latency and reliability are very stringent compared to eMBB, with a latency target of less than 1 ms and reliability equal to 99.9999%.

1.2 Scope and contributions

The coexistence of the URLLC and eMBB slice is a very challenging task. Their multiplexing on the same infrastructure needs to be done in an effective and optimal manner. The management of the radio resource allocation is definitely a crucial, still open research problem. With the emergence of

several technologies that are added to the RAN, the RRM procedures become more sophisticated and complex. In this thesis, we aim to answer the following questions relative to resource management:

- What management mechanisms can be put in place to jointly manage several slices deployed on the same infrastructure?
- How to ensure the differentiated monitoring of the slices and how to exploit measurements to enrich the management system of the RAN?
- Depending on the evolution of standardization, how to evolve the management of the RAN to integrate the new verticals?
- What should be monitored and with what granularity? What post-processing should be implemented in terms of data analysis to detect inconsistencies/dysfunctions?

We aim in our work to answer some of these questions for the case of URLLC and eMBB co-existent slices. Our target is to implement the radio resource management mechanisms necessary to control these different slices. Guaranteeing the SLA of one slice should not degrade the SLA of the other. This isolation between slices cannot be guaranteed by over-dimensioning the reserved resources. As the radio resources are scarce, it is necessary to use them as accurately as possible, without over-dimensioning and without degradation of the performance of the different slices.

We propose two dimensioning models, making use of the radio conditions of the users obtained by the network. A first one that is based on packet queuing and a second one based on large deviation bounds. We dynamically allocate resources between slices with an inter-slice scheduling method that guarantees vehicular URLLC's SLA while limiting eMBB performance degradation. The use-cases are simulated on a 5G system simulator described in 3.2.2.

The manuscript is structured as follows:

- In Chapter 2, we define some of the 5G slicing enablers including slice creation, slicing management and RRM procedures for resource allocation. We also introduce the slice's reconfiguration delay problem in the case of multiple slices' coexistence, which motivates our proactive resource reservation approaches.
- In Chapter 3, we list some of the related works that address the topic of slices multiplexing and RRM procedures in case of slicing. We describe the system model adopted in the thesis along with the developed simu-

lator. We then detail the proactive resource reservation approaches we propose to solve the extra delay problem due to reconfiguration.

- In Chapter 4, we describe the dimensioning model we propose to obtain optimal allocation of resources between slices. This model is based on radio statistics and the SLA of URLLC slice. We apply this model on a 5G network simulator and analyse the gain in performance for URLLC and eMBB slices.
- In Chapter 5, we introduce a dimensioning model based on large deviation bounds. We compare the analytical model with numerical simulations on our 5G system simulator.
- Chapter 6 contains the conclusion and some hints on future work perspectives, both short and longer terms.

1.3 Publications

1. N. Naddeh, S. Ben Jemaa, S-E. El Ayoubi and T. Chahed, "Proactive RAN Resource Reservation for URLLC Vehicular Slice," 2021 IEEE 93rd Vehicular Technology Conference (VTC2021-Spring), 2021, pp. 1-5, doi: 10.1109/VTC2021-Spring51267.2021.9448703. [7]
2. N. Naddeh, S. Ben Jemaa, S. E. Elayoubi and T. Chahed, "Anticipatory Slice Resource Reservation for 5G Vehicular URLLC Based on Radio Statistics," 2022 IEEE 33rd Annual International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC), 2022, pp. 1019-1025, doi: 10.1109/PIMRC54779.2022.9977792. [8]
3. S-E. El Ayoubi, N. Naddeh, T. Chahed and S. Ben Jemaa, "A Large Deviations Model for Latency Outage for URLLC", EAI VALUETOOLS 2022 - 15th EAI International Conference on Performance Evaluation Methodologies and Tools, November 2022, virtual conference.

Chapter 2

URLLC slicing enablers in 5G RAN

In this chapter, we describe the 5G-NR features that enable the management of the 5G use cases described in Chapter 1 through network slicing. We first focus on the architecture aspects, including the 5G overall architecture, the slice instance generation and the management framework for slices. We then move to the slicing enablers from radio perspective that enable a slice-specific radio resource configuration and allocation. We finally expose the radio resource reconfiguration delay problem and its expected impact on the vehicular URLLC slice performance.

2.1 Architectural enablers for slicing

2.1.1 Overall 5G architecture

The Fourth Generation (4G) and LTE were a breakthrough in terms of higher data rates and the adoption of packet-only switching technique along with the overall change in the core network. The 4G network architecture is illustrated in Figure 2.1 based of 3GPP standards [9]. The User Equipment (UE) is connected to the Evolved-UMTS Terrestrial Radio Access Network (E-UTRAN), which is composed of several base stations called eNodeB. The core of the LTE network contains several entities as follows:

- The Mobility Management Entity (MME) manages UE access network and mobility, paging and bearer path establishment. The Home Subscriber Server (HSS) is the master database to which the MME sends

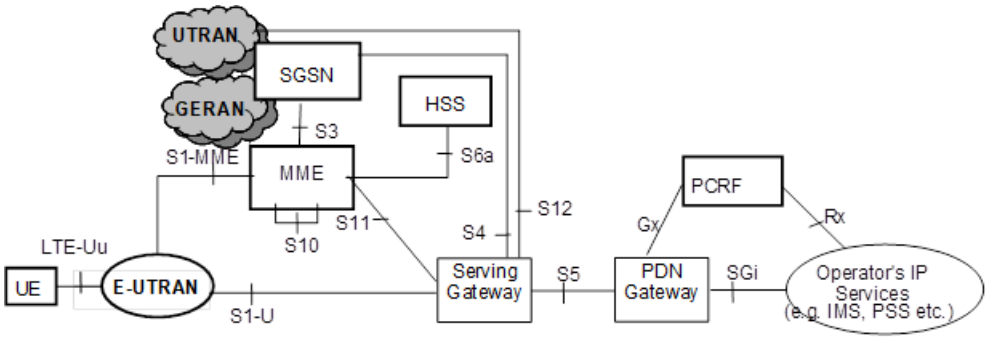

Figure 2.1: 4G architecture [9]

authentication of a UE.

- The Serving Gateway (SGW) forwards and routes user data packets and manages inter-eNodeB mobility.
- The Serving GPRS Support Node (SGSN) provides packet switching, authentication and security.
- The Packet Data Network (PDN) or PGW is the connecting point between the UE and the external network, where several functions are applied on the packets, including screening and filtering.
- Finally the Policy and Charging Rule Function (PCRF) controls the service quality by ensuring good packet flow and policy enforcement.

The traffic growth and the increase in QoS requirements made the 4G insufficient in terms of data rates, latency, and flexibility. Therefore, finding a solution that fits this diversity was necessary. 5G networks make use of the separation of the User Plane (UP) and Control Plane (CP) functions that allow having both centralized and distributed resource allocation schemes, an interaction between the network virtual functions, higher data rates, lower latencies, higher flexibility, and scalability. 5G provides a much more flexible RAN compared to the previous 3GPP RAN, with the integration of Software-Defined Networking (SDN) and Network Function Virtualization (NFV), to meet the diverse requirements of different services. 5G RAN supports multiple air interfaces, for example, below 6 GHz and mmWave spectrum bands, and tight integration with other licensed spectrum and unlicensed spectrum Radio Access Technologiess (RATs) with their unique capabilities. This diversity would be critical for enterprise and industrial environments that deploy other networking technologies like time-sensitive networks for their specific use cases. Tight integration with 5G would support service continu-

ity among different networks and allows wide-area coverage access even for localized services of the industrial and enterprise tenants.

The 5G architecture for Release 15 depends on the 4G network architecture. It is called the Non-Standalone Fifth Generation-New Radio (NS 5G-NR), which means the existing 4G infrastructure will support the 5G networks. Figure 2.2 illustrates a detailed architecture of the 5G system proposed by the 3GPP.

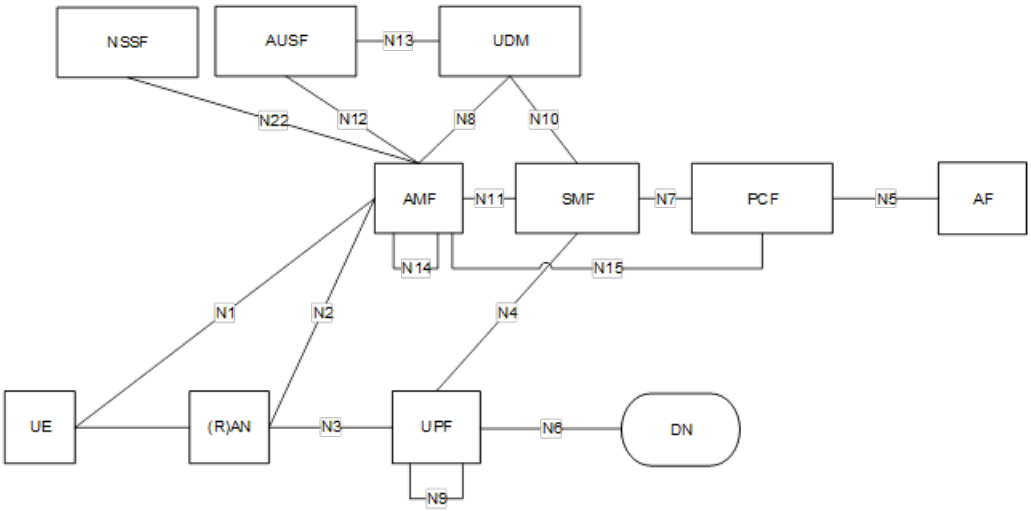

Figure 2.2: Non-roaming 5G System architecture from 3GPP Release 15 [10]

We cite the Network Functions and Entities illustrated in Figure 2.2:

- Access and Mobility Management Function (AMF): UE-based authentication, authorization, and mobility management
- Authentication Server Function (AUSF): UE authentication data
- Data Network (DN): Identifies Service Provider services
- Unstructured Data Storage Function (UDSF): master database to store dynamic data
- Network Exposure Function (NEF): manages the external open network data
- NF Repository Function (NRF): allows NF to register and discover each other

- Network Slice Selection Function (NSSF): Select the Network Slice Instance (NSI), as will be detailed next
- Policy Control Function (PCF): Policy control and QoS
- Session Management Function (SMF): UE session management and IP address allocation
- Unified Data Management (UDM): UE subscription data management
- Unified Data Repository (UDR): converged repository to store data
- User Plane Function (UPF): UE data transfer
- Application Function (AF): Application relocation or reselection and PCF policies adjustment
- 5G-Equipment Identity Register (5G-EIR): Independent network component coupled via Service Based Interfaces (SBI) that helps telecom operators protect their networks
- Security Edge Protection Proxy (SEPP): Ensures end-to-end confidentiality and/or integrity between source and destination network

We now provide a general overview of the 5G RAN architecture and its interaction with the 5G Core (5GC) (Figure 2.3).

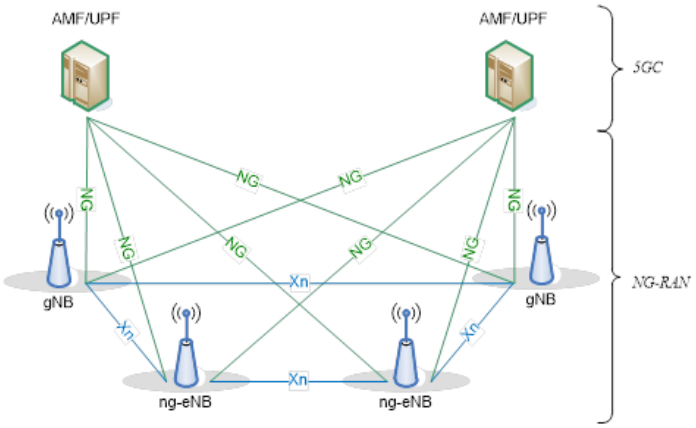

Figure 2.3: Overall RAN Architecture [11]

An NG-RAN consists of nodes. It is either:

- a Next Generation Node B (gNodeB), providing New Radio (NR) UP and CP protocol terminations towards the UE

- an next generation enhanced Node B (ng-eNB), providing E-UTRAN UP and CP protocol terminations towards the UE

They are interconnected with an Xn-interface and connected to the 5GC with Network Getaway (NG) interfaces, such as Network Getaway User (NG-C) to AMF and Network Getaway Control (NG-U) to UPF.

2.1.2 Management framework for network slicing

3GPP has defined several network management services related to network slicing operations. These management services aim to provide effective means for both intra-slice management, which deals with the management of individual network slice, and inter-slice management, which deals with the management of multiple simultaneously operational network slices. These management services integrate into the overall 5G system, which can be broadly categorized into three different layers as depicted in Figure 2.4.

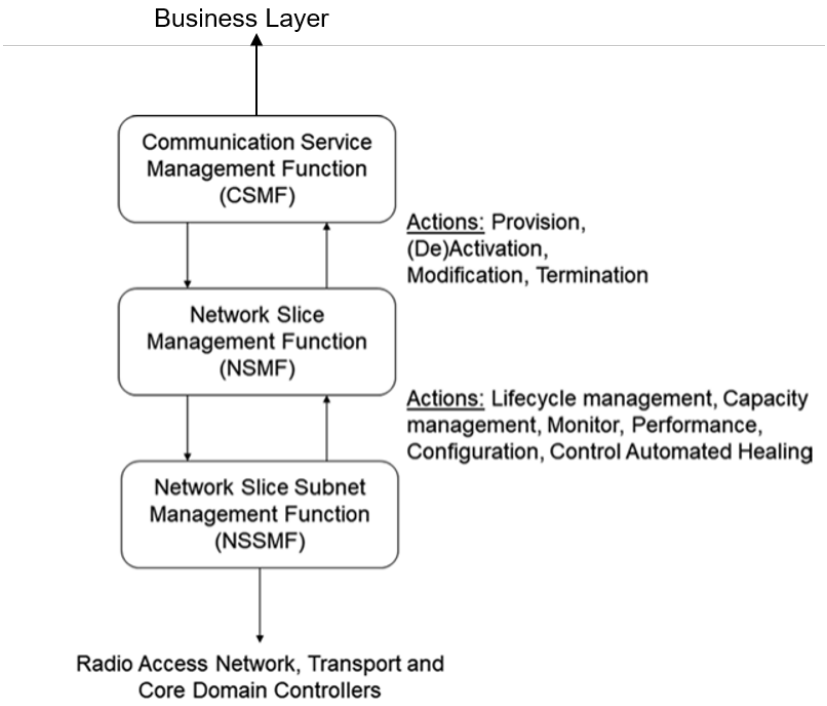

Figure 2.4: Management Framework [12]

The business layer provides the network Business Support Systems (BSS) required for the business-level policy definition and implementation. This

layer also provides the interfaces and applications for interaction with external entities such as the network slice tenant or the external infrastructure providers for network deployments with network slicing. For example, the network slice tenant can use the provided Application Programming Interfaces (API) to request a network slice or change existing network slice provisioning.

We focus in the Figure 2.4 on the Network Slicing Management layer which we now describe.

Network Management functions

These functions provides the network slicing management functionalities. the following management functions have been defined with their scope of responsibilities related to network slice-specific management.

- Communication Service Management Function (CSMF) converts the business-level service requirements given by the business layer into network slice-specific technical requirements.
- Network Slice Management Function (NSMF) is responsible for the E2E management of one or more NSI. It create the NSI life-cycle management, stores NSIs mapping relationship in a storage base, and implements the NSI management functions. NSMF converts the E2E service level requirements, as provided by the CSMF, into domain-specific requirements like RAN, Core, and Transport network. When the NSMF manages multiple NSIs, it might also perform some inter-slice coordination in resource allocation and management function isolation among different network slices.
- Network Slice Subnet Management Function (NSSMF) performs the management of one single domain like RAN or Core Network. The domain-specific management is governed by the E2E slice-level requirements and may use the management services of other management functions, for example, related to the individual network elements. If multiple slices share a network slice subnet, NSSMF must ensure proper resource allocation and isolation among those slices as required by the NSMF.

The slice specific information and orchestration given by the network management functions are eventually given dependent of their type and domain to the RAN, Core and transport domain controllers.

2.1.3 Network slicing in the RAN

Network slices aim to ensure service guarantee from E2E perspective, which includes the RAN performance. E2E service requirements, therefore, are mapped to corresponding RAN level requirements. It includes the requirements of the individual network slice and the coordination and prioritization policies for the simultaneous operation of multiple network slices on the same RAN elements and functions.

Especially in cases where network slices have extremely divergent requirements like eMBB and URLLC, careful planning needs to be done regarding the deployment and configuration of different RAN elements and functions. For network slicing support in RAN, 3GPP defined several basic principles, among which [11]:

- RAN Awareness of Network Slices and QoS Differentiation: 5G RAN should support configuration and handling for a differentiated traffic processing for different slices. 3GPP has specified numerous physical layer features that allow a flexible configuration of the shared physical layer for different slices. Also , the NG-RAN should select the RAN part of the network slice that is slice identification provided by the UE or the 5GC.

- Resource Management Between Network Slices: 5G RAN network elements may support multiple network slices, which requires proper RRM policy enforcement. The main objectives of such RRM policies are to ensure: 1) fair resource utilization according to the SLA with the tenant, and 2) proper resource isolation to ensure that congestion in one slice does not affect the performance of another slice.

- Network Slice Geographical Availability: A network slice may be available in the whole network operator coverage area or only a part. A network slice with limited coverage support should not be accessible outside of this specified geographical area.

- UE Association with Network Slices: A UE may be authorized to access many network slices, but it can be simultaneously associated to a maximum of eight network slices. To enable the UEs and Network Elements/Functions to identify different network slices a parameter called Network Slice Selection Assistance Information (NSSAI) is introduced [10] described next.

2.1.4 Network slice instance

A NSI is defined within a Public Land Mobile Network (PLMN). An Single-Network Slice Selection Assistance Information (S-NSSAI) defines a Network Slice. It is composed of the following:

- a Slice/Service Type (SST) that defines the features and services of a slice. Three values for SST have been standardized to represent general service types of eMBB, mMTC, and URLLC, but operators also have the possibility to define their own values for a much finer categorization of supported network slices
- a Slice Differentiator (SD) that is optional to differentiate multiple NS having the same SST

The NSSAI is a collection of S-NSSAIs. An NSSAI may be a Configured NSSAI, a Requested NSSAI, or an Allowed NSSAI. There can be at most eight S-NSSAIs in Allowed and Requested NSSAIs sent in signaling messages between the UE and the Network. Based on the Requested NSSAI and the Subscription Information, the 5GC is responsible for selecting a NSI(s) to serve a UE, including the 5GC CP and UP Network Function.

2.2 Radio Resource Management for multiple slices

2.2.1 Inter- and intra-slice scheduling

Network slicing management is one of the most challenging tasks in 5G-NR. The 5G RRM is based on two levels of scheduling, the inter-slice, and intra-slice resource allocation.

At the first level of inter-slice, the scheduler uses the RAN slicing technique to manage 5G radio capacity for different services. This scheduling needs to be intelligent, flexible, and cognitive to meet the diverse QoS requirements of 5G streams. At the second level of intra-slice allocation, the scheduler allocates the available resources among users attached to a specific slice. The intra-slice scheduler makes use of classical scheduling schemes (e.g., round robin, best Channel Quality Indicator (CQI), or proportional fair). It applies to users of the same type in each 5G slice.

The inter-slice resource allocation resolves the issue of scheduling radio resources in the RAN given a specific SLA for a slice. The SLAs target a variety

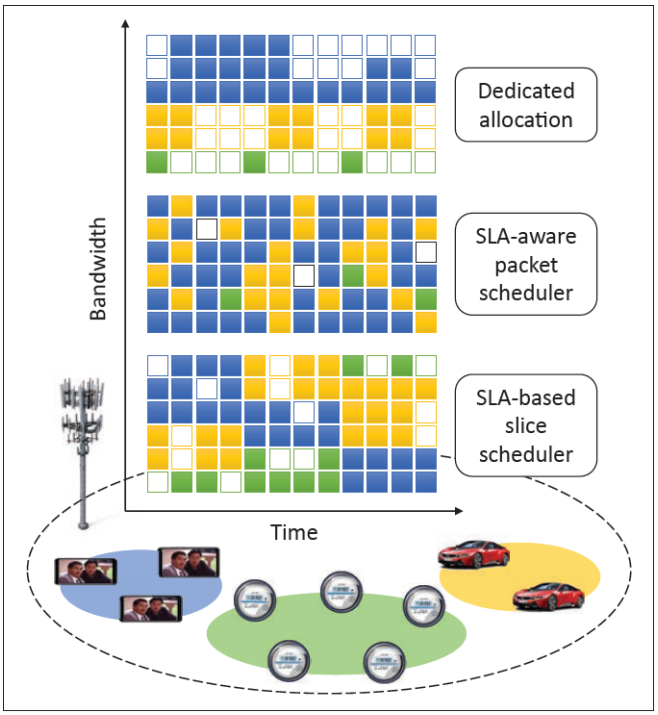

Figure 2.5: Inter-slice radio resource allocation schemes. [13]

of Key Performance Indicators (KPIs), such as throughput, latency, reliability, and availability. We apply this allocation at different time granularity, depending on the types of slices and demands in the RAN. As mentioned before, the slices share and reuse the physical infrastructure of the RAN, which is the main idea of inter-slice. We classify the existing schemes of Inter-slice resource allocation into three categories: the dedicated allocation, the SLA-aware packet scheduler, and the SLA-based slice scheduler, which we illustrate in Fig.2.5.

The dedicated allocation provides each slice with a part of the spectrum after its creation and then keeps this assignment unchanged during the entire life cycle. This allocation has the highest isolation level for slices and allows high-level scheduling for each slice. Life-critical services are the most favorable for such allocation. However, we must allocate large amounts of resources to meet each slice's SLA. This dedicated allocation may cause inefficiency in resource utilization, especially with dynamic services.

The SLA-aware packet scheduler is to have a dedicated scheduler for all slices. This scheduler can guarantee each user's QoS and the slice's SLA since it has flexibility in the time and frequency domain. However, the complexity

of having diverse QoS and SLA makes this solution suitable for slices of similar requirements.

The SLA-based slice scheduler is a trade-off between the previous schedulers. We can allocate resources periodically (in a given window time) between slices dynamically, with a target of satisfying slices' SLA. The scheduler applies the dedicated scheduler or a softer solution of assigning a number of radio resources per slice. So we can apply an independent scheduler on each slice, or a low complex two-level scheduler can be introduced to the existing ones to help lower multiplexing complexity. The scheduling applied uses different time windows depending on the level of isolation of the slice (higher isolation requires a bigger window and vice-versa).

2.2.2 Numerology and frame structure

The new Numerology in 5G-NR aims to help slices achieve their KPIs, specific latency for URLLC and data rates for eMBB. Numerology is defined by Sub-Carrier Spacing (SCS) and Cyclic Prefix (CP) overhead. Adjustable SCS for different slot duration is essential in realizing QoS in diverse services. 5G technology [14] supports five types of sub-carrier spacing depending upon the numerology type as mentioned in Table 2.1.

The scalability in the 5G numerology helps critical verticals with low latency achieve their target. This is due to the different granularities in the time axis.

Numerology (μ)	SCS ($2^{\mu}\cdot 15kHz$)	slot duration(ms) (14 symbols)
0	15	1
1	30	0.5
2	60	0.25
3	120	0.125
4	240	0.0625

Table 2.1: 5G Numerology

A slot is based on 14 Orthogonal Frequency Division Multiplex (OFDM) symbols and is transmitted within a Transmission Time Interval (TTI). The variable numerologies and the need for smaller transmission time created what we call a mini-slot or sTTI. A mini-slot in NR can start at any OFDM symbol and can be of a length of 2, 4, or 7 symbols as defined in the standard [15]. This provides fast transmission for URLLC. Thus, mini-slots is the ideal solution to low-latency transmissions disregarding of SCS.(Table 2.2)

SCS	Slot	Mini-slot		
	(14 symbols)	(7 symbols)	(4 symbols)	(2 symbols)
15 kHz	1 ms	0.5 ms	0.286 ms	0.143 ms
30 kHz	0.5 ms	0.25 ms	0.143 ms	0.071 ms
60 kHz	0.25 ms	0.125 ms	0.071 ms	0.036 ms
120 kHz	0.125 ms	0.63ms	0.036 ms	0.018 ms

Table 2.2: Mini-slot duration for different numerologies

As described above, in 5G-NR multiple numerologies are supported, and the Radio Frame (RF) structure gets a little bit different depending on the type of numerology. However, regardless of numerology, the length of one RF and the length of one subframe is the same. The length of a RF is always 10 ms, and the length of a subframe is always 1 ms, which gives us the frame structure in Table 2.3 for the different numerologies.

μ	N_{symb}^{slot}	$N_{slot}^{\text{frame},\mu}$	$N_{slot}^{\text{subframe},\mu}$
0	14	10	1
1	14	20	2
2	14	40	4
3	14	80	8
4	14	160	16
5	14	320	32
6	14	640	64

Table 2.3: NR Radio frame structure [16]

The maximum number of Resource Blocks (RBs) for Down-link (DL) and Up-Link (UL) is defined in [17] and illustrated in Figure 2.4 and 2.5. Following is the maximum number of RBs we can configure in Radio Resource Control (RRC) message and Data Center Interconnect (DCI). In terms of RF, we may need a bit wider bandwidth than this because we need to consider the guard band.

SCS(kHz)	5 Mhz	10 Mhz	15 Mhz	20 Mhz	25 Mhz	30 Mhz	35 Mhz	40 Mhz	45 Mhz	50 Mhz	60 Mhz	70 Mhz	80 Mhz	90 Mhz	100 Mhz
15	25	52	79	106	133	160	188	216	242	270	N/A	N/A	N/A	N/A	N/A
30	11	24	38	51	65	78	92	106	119	133	162	189	217	245	273
60	N/A	11	18	24	31	38	44	51	8	65	79	93	107	121	135

Table 2.4: Maximum transmission bandwidth configuration (1)

SCS(kHz)	50 Mhz	100 Mhz	200 Mhz	400 Mhz	800 Mhz	1600 Mhz	2000 Mhz
60	66	132	264	N/A	N/A	N/A	N/A
120	32	66	132	264	N/A	N/A	N/A
480[1]	N/A	N/A	N/A	66	124	248	N/A
960[1]	N/A	N/A	N/A	33	62	124	148
Note 1 : This SCS is optional in the release [17]							

Table 2.5: Maximum transmission bandwidth configuration (2)

2.3 Challenges for resource allocation for vehicular URLLC and eMBB

2.3.1 Intra-slice resource configuration

As mentioned in chapter 1, the URLLC services are very critical in terms of requirements [18]. The high reliability and low latency characteristics make the coexistence with other slices more challenging. Depending on the use case, for instance, Remote Driving, the maximum E2E latency is set to 5ms and reliability to 99.999% [18] [19].

When it comes to intra-resource management for the URLLC slice, there is a need for specific numerology, with an sTTI and robust channel coding. This numerology is to be selected based on an algorithm that takes as input the average SNR, the Doppler effect, and delay spread [2].

In addition to these general URLLC features, when it comes to vehicular services, specific features for coping with high mobility and combat Doppler effect are also provided, with a larger SCS as described in [20]. 3GPP dedicates one or more specific slices to vehicular services with an appropriately configured numerology in a given bandwidth range.

2.3.2 Inter-slice resource allocation and the delay reconfiguration problem

Using the techniques used in the intra-slice resource allocation requires that resources are always available for URLLC. In this case, latency is only due to the packet alignment, scheduling grant reception, over-the-air transmission and packet decoding. So what happens when resources are insufficient or the traffic load is high? We need to consider an additional component, the queuing delay, i.e., the delay before a resource is available for the packet to be scheduled. This queuing delay must be added to the other delay components and considered in the overall radio and E2E latency.

Vehicular URLLC faces the problem of congestion in a gNodeB. Upon the arrival of a vehicular URLLC to a gNodeB, the NSSMF checks if there are enough resources to be attributed. If so, they are allocated immediately to the slice, as enabled by the seamless handover enhancements of 5G NR [21].

However, suppose resources are unavailable for the new user and those already present in the gNodeB. To adapt to the traffic dynamicity, re-configuring the network slice must be performed adaptively.

If the gNodeB is not configured correctly with the required URLLC slice resources, it will negatively impact URLLC reliability. Preemptive priority is traditionally considered a solution for ensuring URLLC QoS without resource reservation. Preemption is only possible if the resources for the slices are configured with the same numerology (Sub-Carrier Spacing, Cyclic Prefix, channel access) but a different mini-slot size. For services that require specific numerology, preemption is not possible. Therefore RRC re-configuration [22] and Bandwidth Part (BWP) reconfiguration [23] are required before eMBB resources can be reused by URLLC. [24] observed that BWP reconfiguration for a UE may take up to 80 ms.

In the remainder of this thesis, we will present algorithms and models for solving the following two main problems related to vehicular URLLC slice resource allocation:

1. Resource dimensioning for URLLC slice: we will provide numerical and analytical models for computing the amount of resources to allocate for the URLLC slice, knowing the traffic load and some statistics on the radio conditions.
2. Resource reservation schemes for URLLC slice: We will propose proactive resource reservation schemes for coping with the reconfiguration delay problem, when preempting resources from the eMBB slice to the URLLC one.

Chapter 3

Proactive versus reactive resource allocation

In the previous chapter, we introduced the notion of network slicing and evoked how slice management and frame structure can help optimize resource allocation and performance. However, for vehicular applications, reconfiguration the slices in the new cell may introduce delay that may have a negative impact on URLLC performance. This chapter assesses, by simulations, the impact of the reconfiguration delay on the performance and proposes a proactive scheme that anticipates the vehicle arrival for slice reconfiguration. We first start by a literature review on resource allocation schemes for URLLC slices.

3.1 Related works

Many papers in the literature deal with the RRM mechanisms that allow reaching low latency on the radio interface when multiplexing URLLC with eMBB, especially RAN slicing configuration. In [25], authors propose different options for configuring RAN slices using a set of control parameters that dictate the operation of the packet scheduling function at Layer 2 and the Radio Admission Control (RAC) function at Layer 3. They evaluate the impact of these parameters on having an efficient radio resource by changing parameters such as bandwidth, priority, or SCS. Their results show that L3 parameters impact slice isolation cases more than L2 parameters. The work of Feng *et al.* in [26] proposes a dynamic resource allocation scheme for eMBB and URLLC slices. This scheme is based on optimal power control for latency-aware resource allocation. The results show significant results in

achieving low latency, but dynamic bandwidth allocation was not discussed in the paper. In [27], authors proposed a dynamic Bandwidth Slicing and Resource Allocation (BSRA) framework for IoT and video streaming slices in a virtualized network. They applied this framework on a long timescale and IoT scheduling and power control on a short time scale. They aim to minimize the total cost by applying the Lyapunov optimization method. Their results show that compared to static bandwidth slicing, the performance of both services is improved when it comes to power-delay and cost-delay trade-offs. Although it gives good results, it does not explicitly target URLLC reliability and packet latency.

Authors in [28] propose an inter-slice scheduler based on multi-object Markov Decision Process (MDP). It allocates resources efficiently between eMBB and URLLC slices on shared bandwidth. They use Probabilistic model checking to analyze the scheduler's performance and perform quantitative verification. They use Pareto curve regions to get strategy synthesis and model zone. The work in [29] proposes a new RRM mechanism and compares it to the existing ones, showing their shortcomings. They defined a new system model for multiple slices and studied the impact of slice-specific control parameters on KPIs. The authors in [30] proposed slice-aware resource allocation for multiplexing eMBB and URLLC with service isolation. They aim to maximize the sum rate of the network, by formulating an AMC resource optimization algorithm. The RB allocation and link adaptation are based on the SINR and consider the MCS selection in the design of the resource allocation algorithm. Although this algorithm has a high average sum-rate for eMBB, it fails in providing the QoS of the URLLC in terms of latency and targets a Block Error Rate (BLER) of 0.001, which does not meet the URLLC reliability constraint. In [31], the authors propose a distributed Machine Learning (ML) solution for proactive RRM in case of scheduled and non-scheduled URLLC. Authors in [32] aim to maximize eMBB throughput while ensuring URLLC latency. They propose a dynamic programming approach that optimizes resource allocation and applies it on top of heuristic scheduling algorithms. The work in [33] designed a scheduling policy that aims to optimize the reliable latency performance of a URLLC user in a Multiple-Input-Single-Output (MISO) system under statistical delay constraints.

In addition, a large number of papers in the literature deal with the new features added, allowing the reach of low latency and high data rates on the radio interface when multiplexing URLLC with eMBB. Authors in [34, 35] examine the impact of changing the TTI length dynamically on serving the URLLC packets while meeting the deadline and guarantying eMBB per-

formance. Other works discuss the Semi-Persistent Scheduling (SPS) approach [36]. [37] computed the amount of resources reserved for URLLC users, knowing a deterministic traffic pattern and target reliability. In [38], a semi-persistent DL scheduler is proposed to pre-allocate resources based on a short-term prediction of arriving traffic. They apply predictive user priority functions to the scheduler to maximize the total throughput of the network and throughput fairness. The resource allocation algorithm presented in [39] targets maximizing resource utilization and increasing eMBB throughput but fails to attain low V2X latency.

As of URLLC intra-slice resource allocation, many works proposed grant-free contention-based channel access for URLLC in the UL. Authors in [40] proposed to send these replicas in a contention-based manner on different frequency resources on consecutive time slots. In contrast, in [41], the authors considered a more flexible scheme where replicas can be sent on any of the available time-frequency resources. These schemes focused on the UL, as the centralized orthogonal resource allocation in the DL is supposed to avoid collisions between packets. However, in high-traffic regimes, the problem of resource dimensioning is still open, even for the downlink. Fro example, in some industrial or vehicular situations, URLLC traffic load may be large, and the (local) network operators must provide sufficient resources while avoiding over-dimensioning.

When dealing with inter-slice resource allocation, preemptive scheduling has been studied in several studies. Authors in [42] present a new scheduling algorithm where URLLC traffic is dynamically multiplexed through puncturing the eMBB traffic, with an added recovery mechanism for punctured eMBB packets. In [43], a new scheduling algorithm is presented, allowing a joint eMBB and URLLC scheduling process. URLLC and eMBB users are multiplexed on the same bandwidth using puncturing. In [44], the authors propose a joint optimization framework for URLLC and eMBB with preemptive scheduling to achieve better URLLC performance while limiting the impact on eMBB throughput.

Priority scheduling has also been the subject of a large pan of the literature. The authors in [45] propose a priority-based resource reservation mechanism to decrease URLLC delay and increase reliability. However, their solution does not reach the latency constraint for a URLLC. [27] propose a bandwidth slicing algorithm for multiple services for a virtualized network, but they did not consider the impact of critical services on the performance of eMBB service.

The objective of the previous papers was to achieve flexibility for serv-

ing URLLC in the presence of lower-priority eMBB services, always with the assumption of a sufficiently large amount of resources in the cell and an eMBB slice configuration that allows serving URLLC on eMBB resources with acceptable performance. We will propose in this thesis accurate resource dimensioning methods for URLLC resources and proactive slice reconfiguration schemes for coping with the delay problem. We will first assess, by simulation, the impact of the reconfiguration delay.

3.2 Assessing the impact of reconfiguration delay

3.2.1 System model

The network consists of a set of K gNodeBs; each one supports two slices for eMBB and vehicular URLLC services [46]. Spectral resources should be distributed among these slices to satisfy URLLC SLA - a packet loss on the order of 10^{-5} and a radio delay[1] less than 3ms- while minimizing the degradation of eMBB throughput.

Network slices are managed at the RAN level by the NSSMF, where most of the slice intelligence resides [47]. In particular, the NSSMF should collect and store essential slice quality information, such as observed CQI distribution and QoS, and make intelligent decisions about resource reservation and scheduling schemes.

We now determine the number of RBS needed for carrying a URLLC packet. We consider a gNodeB where the URLLC slice has to carry packets that belong to different users and thus use different Modulation and Coding Scheme (MCS) depending on the calculated Signal to Interference Noise Ratio (SINR).

For an MCS i, let the spectral efficiency be equal to e_i (bit/s/Hz). For an application packet of size a bits, a bandwidth per RB of bHz and a sTTI length of T, the number of physical RBs, R_i, for transmitting an application packet with MCS i is given by:

$$R_i = \lceil \frac{a}{e_i T b} \rceil \tag{3.1}$$

$\lceil x \rceil$ being the smallest integer larger than or equal to x. While a depends

[1] We consider here the radio delay, which is a component of the E2E delay. For Vehicular URLLC service, the E2E maximum delay is typically set between 5 and 20 ms [4]

on the application, b and T depend on the radio configuration, and e_i on the chosen MCS.

With a total available spectrum for URLLC transmission of B_u (in Hz), the total number of RBs is equal to B_u/b (usually an integer), and the number of URLLC packets that can be multiplexed per slot is obtained from equation (3.1) by:

$$K_u(B_u) = \lfloor \frac{B_u/b}{R} \rfloor = \lfloor \frac{B_u/b}{\lceil a/(eTb) \rceil} \rfloor \tag{3.2}$$

where R is the number of RBs per packet, knowing that the considered MCS has a spectral efficiency of e. $\lfloor x \rfloor$ is the largest integer smaller than or equal to x.

The resource allocation for each packet is done with two approaches:

- Non-flexible allocation: The packets are served on the available resources one by one. When the resources are not enough for a packet to be served, the whole packet is then queued to the next TTI and the remained resources are lost.
- Flexible allocation: In this case, if the resources are not enough to serve a whole packet, the packet is scheduled on two consecutive TTIs without loosing resources.

As of the eMBB performance, let the total available spectrum in the gNodeB be equal to B Hz and the reserved bandwidth for URLLC users at time t be given by $B_u(t)$. The remaining resources $B - B_u(t)$ are shared among the active eMBB users. At time t, let $n(t)$ be the number of active eMBB users, and $e_i(t)$ be the spectral efficiency of the MCS selected by eMBB user i, the instantaneous throughput for user i is then given by:

$$T_i(t) = \frac{(B - B_u(t))e_i(t)}{n(t)} \tag{3.3}$$

3.2.2 Simulator description

We implement a 5G network with RAN slicing in a C++ simulator. We illustrate the simulator's general block diagram in Figure 3.1. First we initialize the slices, their SLAs, the UE classes, the network map, then each time step:

1. the UEs move
2. the radio conditions are updated and reconfiguration is made accordingly

3. the communication procedure gets updated
4. the UEs (in the Receive Data procedure) get scheduled and receive data accordingly
5. the UEs that finished the download (or cannot reconnect to a gNodeB) are deleted for eMBB slice, URLLC users remain in the network
6. new users are generated randomly depending on their classes.

After assigning the bandwidth B_u and B_e to URLLC and eMBB slices, respectively, and their RBs RB_u and RB_e, we schedule the users separately. The packets of the URLLC users are queued in a First-Come-First-Served (FCFS) queue. We detail in Algorithm 1 the scheduling steps for URLLC packets in one sTTI in a gNodeB.

B_u and T being the BWP allocated to URLLC slice and sTTI, respectively, as mentioned before. $bits\_per\_RE$ is the spectral efficiency for a given SINR. $current\_sTTI$ and $sTTI\_arrival$ are the time slot of the transmission and the arrival of a packet, respectively. D_u and O_u are the packet delay and the gnodeB outage respectively. Using these parameters, we calculate the overall average packet loss and delay for each gNodeB.

Finally the KPIs are extracted for performance analysis. Note that in this work, the simulator's granularity is of the order of 100 ms. We reset the measurements at the beginning of each time step, and update them throughout each time step.

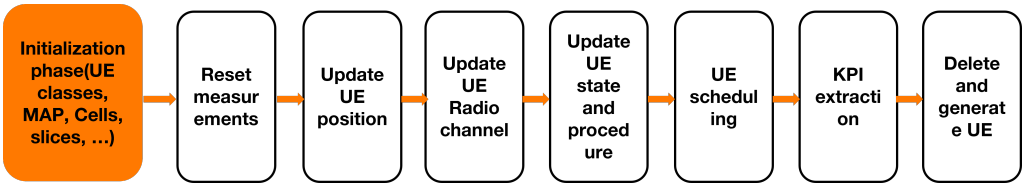

Figure 3.1: 5G-NR Simulator block diagram

3.2.3 Reconfiguration impact illustration

In order to capture the impact of the slice reconfiguration delay on the URLLC performance, we simulate a 5G network composed of 13 gNodeBs forming a three-sectored deployment with *500* meters inter-site distance, in compliance with the 3GPP urban macro deployment [48], with 20 MHz bandwidth. We implement network slicing in the NSSMF entity for all gNodeBs.

Algorithm 1 URLLC scheduler

Input: T, RB_u
Output: D_u, O_u

remainRB= RB_u
$sTTI\_length = T$
Create packets following Poisson distribution

$$y = poissrnd(\lambda * TTI\_length) \tag{3.4}$$

Calculate the SINR of each UE and affect the corresponding $bits\_per\_RE$ to the packets
for $j \leftarrow 1$ to y **do**
  $RE\_per\_packet = packet\_size/bits\_per\_RE$
  if $remainRB \geq RE\_per\_packet$ **then**
    $D_u = current\_sTTI - sTTI\_arrival$
    if $D_u \leq delay\_max$ **then**
      Transmit packet
      $ok\_packet+ = 1$
    else
      packet lost
      $outage\_packet+ = 1$
    end if
    remainRB$- = RE_per_packet/24$
  else
    loose the remaining capacity and we jump to next sTTI
  end if
  $O_u = numb\_outage\_packet/(outage\_packet + ok\_packet)$
end for

Each gNodeB has two slices: URLLC and eMBB. The slice is created with the following properties: SST [10], label, number of connected users, radio resource percentage, maximum delay, and average throughput. Figure **??** illustrates the network created by the simulator, showing eMBB and URLLC UEs and URLLC vehicle trajectory.

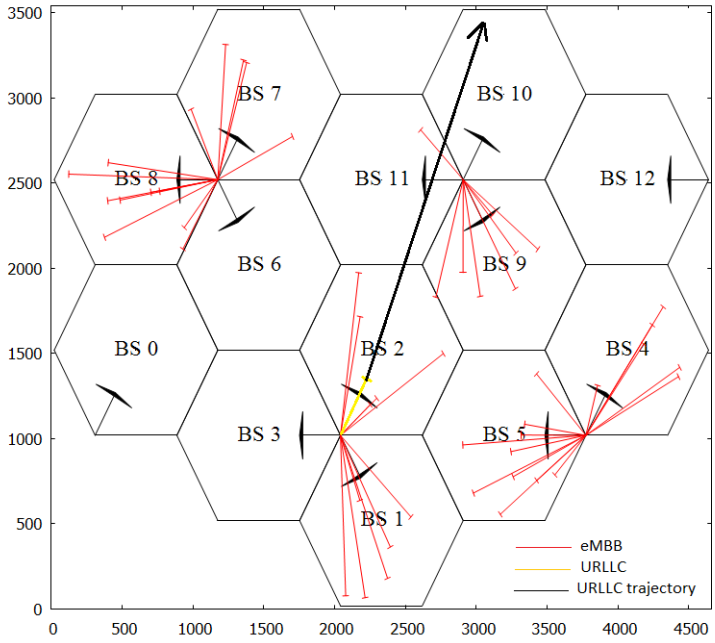

Figure 3.2: Urban network with 13 gNodeBs.

The eMBB users arrive in the network following a spatial Poisson process of mean 3.42 [user/sec/gNodeB]. We consider a File Transfer Protocol (FTP) like traffic of fixed file size, 14 Mbits. Once the file is transmitted, the eMBB user leaves the network. URLLC vehicles are created on the roads following a linear Poisson process with different arrival rates depending on the scenario with a mean of 0.395 [Vehicle/sec/km] and move at an average velocity of 50km/h for a total distance of 2.526km. For each vehicle, small URLLC packets of size *96* bits are generated following a Poisson distribution with mean 2 [packets/msec/vehicle]. The vehicles remain active during the simulation time until they leave the network.

The vehicular URLLC has the following SLA requirements: 10^{-5} of reliability and $5 - 20$ms of E2E latency, which corresponds to radio and backhaul/backbone latency. So depending on the networks' architecture and the services, the operator can choose the radio latency limit. In our case, we

limit the queuing time of radio latency to $1ms$, after which the packet is considered lost.

The simulation and configuration parameters are presented in Table 3.1 [49] [16].

Table 3.1: System parameters.

Parameters	URLLC	eMBB
Environment	3GPP Urban Macro (UMa)	
Number of gNodeBs	13	
Bandwidth	20 Mhz	
Sub-Carrier-Spacing (SCS)	30 Khz	15 Khz
Number of RBs	51	106
TTI size(ms)	0.143	1
Traffic model	Poisson	
Packet size	96 bits	14Mbits
Speed	50 Km/h	Static
Scheduling granularity	sTTI	TTI

We illustrate in Figure 3.3 the vehicular URLLC packet loss during the simulation time. In this simulation, we allocate minimal resources for URLLC and increase the reservation when new vehicles join the gNodeB. With the vehicle's mobility leading to handovers between gNodeBs, we observe peaks of packet losses due to the reconfiguration delay. These peaks vanish for a while until another handover occurs. These peaks can attain a loss of more than 10^{-2}, which is unacceptable for V2X URLLC services. This degradation increases when the intensity of traffic increases, as the minimal amount of allocated resources becomes insufficient in most gNodeBs, which motivates the need for our anticipatory reservation proposals.

3.3 Proactive resource reservation

Based on the observation of degraded URLLC performance due to the slice reconfiguration delay, we propose three proactive resource reservation schemes and compare their performances with the reactive scheme assessed before.

3.3.1 Proposed schemes

Proactive reservation on neighboring gNodeBs Without prior knowledge of the user's trajectory, we suppose in this scheme that when a URLLC

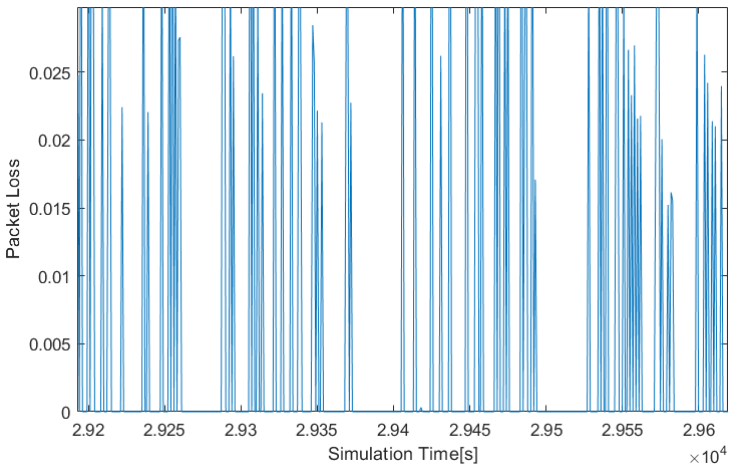

Figure 3.3: URLLC packet loss illustration for reactive resource allocation.

user arrives in a gNodeB, he can move to any of the neighboring gNodeBs. Hence, a corresponding resource reservation is performed on neighboring gNodeBs so that the QoS is guaranteed wherever the URLLC vehicle moves. We describe the steps taken in Algorithm 2:

Algorithm 2 Neighbors anticipated reservation

```
1: if Handover for UE i from gNodeB S to gNodeB T is true then
2:   for all neighbor k of T do
3:     if k does not exist in old neighbor list of S then
4:       Increase N in k
5:     end if
6:   end for
7:   for all neighbor j of S do
8:     if j does not exist in new neighbor list of T then
9:       Decrease N in j
10:    end if
11:  end for
12: end if
```

We consider S the source gNodeB, T the target gNodeB and N the total number of vehicular URLLC UEs in a gNodeB . This scenario enforces URLLC reliability with a possible negative impact on eMBB throughput since we over-reserve the URLLC slice for all neighbor gNodeBs.

Proactive reservation on predicted vehicular URLLC UEs trajectory In this scenario, we suppose that the URLLC UEs' trajectory can be predicted. We can deduce for each gNodeB the expected total number of URLLC UEs and determine the corresponding resource reservation based on the offline study. This procedure is described in Algorithm 3 where we denote by S the source gNodeB, T the target gNodeB, and X the destination gNodeB that follows gNodeB T.

This approach helps us prevent useless reservation of resources and diminishes the impact on eMBB user performance.

Algorithm 3 Trajectory dependent reservation

```
Create User i in gNodeB S
Increase by 1 in next destination T
while User life cycle not equal to 0 do
  if Handover happens then
    Check next destination X to T
    Increase N in X
    Decrease N in S
  end if
end while
```

3.3.2 Baseline schemes

We compare these proactive schemes to two other schemes:

- Static maximal reservation: the resource reservation does not consider the traffic's localization. It corresponds to a classical scenario where higher-level information is not exploited in the lower-level resource allocation. In this case, to target URLLC reliability, a maximal amount of resources is reserved in a static, permanent manner for URLLC slice in all the gNodeBs of the network. This static reservation should have obviously a negative impact on eMBB throughput due to over-reservation; it also corresponds to the extreme case of proactive reservation in space and time.

- Reactive allocation: No pre-reservation is applied, and the reconfiguration happens after the handover.

3.3.3 Offline optimization of the resource reservation

We perform offline simulations on the gNodeBs with a fixed amount of (static) URLLC users in each gNodeB. These simulations allow us to determine the required amount of resources to be reserved to reach the reliability target. The amount of resources reserved for the URLLC slice in each gNodeB is then gradually increased until reaching the target QoS depending on the number of users (the proportion of packets whose delay exceeds 1 ms is equal to 10^{-6}).

The obtained results indicate that for eight active users in the gNodeB, 42% of the gNodeB resources have to be reserved for URLLC. This level increases to 57% for 15 users.

3.3.4 Simulation results

This section applies the previously introduced scenarios with a reactive case on the 5G simulator. We compare the performance of the slices in terms of URLLC reliability and eMBB throughput to assess the impact of proactive reservation on both KPIs.

In Figure 3.4, we illustrate the packet loss for each simulation after reaching a steady state. When comparing the reservation with 80ms reconfiguration to the rest, we observe high packet loss due to reconfiguration delay, while the other three scenarios attain the requested packet loss on the order of 10^{-5}.

This reliability performance has an impact on eMBB throughput. As a result of over-reservation in the static and neighbors reservation scenario, we can see in Figure 3.5 an expected degraded eMBB performance compared to the baseline, where resources are reserved only when and where needed. The unnecessary reservations in gNodeBs where there is a lower number, or no URLLC users are the reason for this impact. When comparing the proactive schemes, we see even though the reservation on neighbors have a slight edge in terms of reliability, however, the impact on eMBB throughput is reduced when the reservation is anticipated only on the trajectory.

3.4 Conclusion

In this chapter, we tackled the problem of slice re-configuration with the existence of multiple slices on a non-shared bandwidth. With this re-configuration resulting in huge packet loss due to latency problems, we proposed proactive

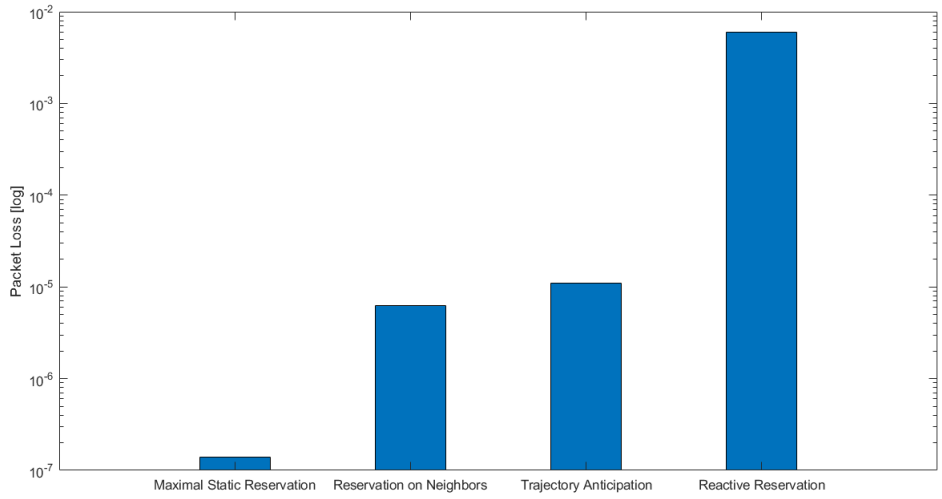

Figure 3.4: URLLC packet loss

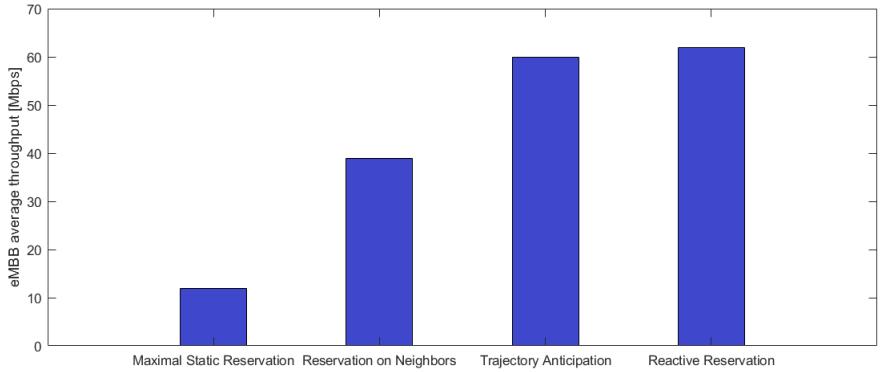

Figure 3.5: eMBB average throughput for all gNodeBs in four scenarios

approaches to solve this issue. We first developed a 5G simulator that implements slices. We simulated the case of a reactive reservation to show the impact of re-configuration on instant packet loss. Along with the proactive approaches, to determine the resource reservation needed to test them, we relied in this chapter on offline simulations that give a good approximation of the number of resources reserved for a given number of URLLC UEs in the gNodeB. However, in a real network, the optimal resource allocation may differ from a gNodeB to another, depending, for instance, on the radio environment, the distribution of URLLC users in the gNodeB, and the quality of transmission experienced by each URLLC user. Hence, this optimal allocation can be determined automatically by the network, which motivates the study of the next chapter, that introduces a different, more optimal way to calculate resources.

Chapter 4

Proactive resource allocation based on radio statistics for vehicular URLLC

In the previous chapter, we introduced the proactive allocation scenarios adopted to respond to the reconfiguration delay issue. The allocation of RBs for each slice depended on offline simulations. These simulations were not optimal since the allocation was not based on the radio conditions of the users. To solve this problem, we introduce in this chapter a dimensioning model based on the knowledge of the MCS distribution of the URLLC users in the different gNodeBs, to calculate the number of RBs allocated per slice. Note that the MCS distribution is computed based on network data statistics and reported to the network management entities. We show that, using this information, the proposed proactive schemes provide the URLLC requirements with low impact on eMBB throughput.

4.1 URLLC performance model

4.1.1 Fixed MCS

We start with a setting where URLLC users always use the same robust MCS that ensures a low BLER. This way, we can avoid additional delays due to channel acquisition, training and MCS adaptation.

URLLC users generate packets sporadically. Let the packet generation process by a URLLC user be Poisson of intensity λ_u packets per second. For

several active URLLC users whose number is equal to N_u, the aggregated packet generation process is Poisson of intensity $N_u\lambda_u$. Packets are generated during a sTTI and wait until the beginning of the next sTTI to be transmitted. Suppose the accumulated number of packets is less than the maximal URLLC capacity $K_u(B_u)$ in the frequency domain, as determined in equation (3.2) (Recall that B_u (in Hz) is the total available spectrum for URLLC transmission). In that case, the remaining packets are stored in a FCFS queue and served in the next time slots.

Let $M_u(m)$ be the number of packets in the URLLC queue at time slot $m \in [0, \infty]$. This number evolves as follows:

$$M_u(m) = M_u(m-1) - \min\left(K_u(B_u), M_u(m-1)\right) + x_u(m) \tag{4.1}$$

where $x_u(m)$ is the number of new packet arrivals during time slot m that is a Poisson random variable of parameter $N_u\lambda_u T$.

For a packet that arrives at time slot m, the worst case radio delay (when it is put at the end of the queue) is computed by:

$$D_u(m, B_u) = T_c + 2kT_{Proc} + (1+2k)T_{tx} + \lfloor \frac{M_u(m)}{K_u(B_u)} \rfloor \tag{4.2}$$

with

$$T_c = 2 * T_{L_1/L_2} + T_a \tag{4.3}$$

where T_{L_1/L_2} is the delay of layer 1/layer 2 processing for eNB and UE, T_a is the delay due to alignment, k is the number of re-transmissions, T_{Proc} is the delay between scheduling request and UL grant, and between DL Hybrid Automatic Repeat reQuest (HARQ) and re-transmission and T_{tx} is the transmission time [50].

This delay is averaged over all time slots that have active user arrivals. The average URLLC delay is given by:

$$\bar{D}_u(B_u) = \lim_{m\to\infty} \frac{\sum_{i=1}^{m} D_u(m, B_u)\mathcal{L}_{x_u(m)>0}}{\sum_{i=1}^{m} \mathcal{L}_{x_u(m)>0}} \tag{4.4}$$

where $\mathcal{L}_c$ is the indicator function that is equal to 1 if condition c is satisfied, and to 0 otherwise.

4.1.2 Heterogeneous MCS

We now consider the case where packets of different users may use different MCS. Let $\mathcal{I}$ be the set of available MCS. When the packet number waiting

to be served equals M_u, the system cannot be completely described by M_u, but also the MCS of each packet. Let $(M_u, \mathbf{I}_u)$ be the system state, with $\mathbf{I} \in \mathcal{I}^{M_u}$ a vector of length M_u, whose k-th element $(I(k))$ is the MCS index for the k-th packet in the queue (the packet at the head of the queue being numbered 1).

Knowing the queue state $(M_u(m), \mathbf{I}_u(m))$ at time slot m, the gNodeB serves in slot m the maximum number of packets $K(m, B_u) \in [1, M_u(m)]$ such that:

$$\sum_{k=1}^{K(m,B_u)} R_{I(k)} \leq \frac{B_u}{b} \tag{4.5}$$

Constraint (4.5) ensures that the consumed resources are limited by the amount of reserved URLLC RBs (B_u/b).

At each time slot, the scheduler serves the first $K(m)$ packets and adds the new arriving packets, whose number is $x_u(m)$ as in equation (4.1), which becomes:

$$M_u(m) = M_u(m-1) - K(m, B_u) + x_u(m) \tag{4.6}$$

The indices of the new packets are also added to $\mathbf{I}_u$.

4.2 Integrating the MCS distribution estimation in the resource allocation framework

In order to compute the amount of resources we need to reserve for URLLC users, we develop a dimensioning module proposed to adjust the reservation based on the URLLC MCS distribution for each gNodeB in the network. When we combine this MCS distribution with the predicted number of URLLC users in a gNodeB, we can estimate the required resource reservation for vehicular URLLC users using the model of section 4.1.2. This scheme can be integrated into a management module in a real network using two approaches:

- Slow dynamics: a centralized management entity takes as input the MCS distribution of the gNodeBs and the estimated traffic during a fixed amount of time. In return, it gives the reservation rate for a specific slice. In this case, the reservation is on the time scale corresponding to the users' mobility dynamics.
- Fast dynamics: a distributed management entity (on each gNodeB) takes as input the MCS distribution of the gNodeBs and the estimated

traffic after each scheduling cycle. The time scale corresponding to this operation goes down to the schedulers' assigned TTI.

In our system model, we choose the slow dynamics approach so that the NSSMF is the management entity that gets radio statistics and gives back the resource reservation per slice. Figure 4.1 illustrates the architecture for implementing the proposed scheme. Within the NSSMF, two modules allow the dynamic management of the slices. First, an MCS distribution module allows for building a per-gNodeB MCS distribution. Second, we use this distribution as input for the resource dimensioning module. This latter takes the estimated traffic (number of URLLC users per gNodeB) and computes the needed amount of resources to be reserved for the URLLC slice in each of the gNodeBs. The system applies the new configuration, dynamically changing depending on the NSSMF updates. These configurations are eventually used to schedule the users.

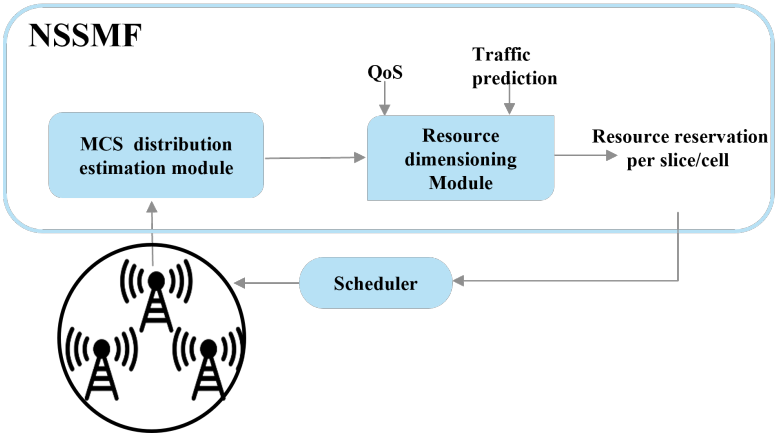

Figure 4.1: Integration of the proposed resource dimensioning module.

MCS distribution estimation module

The first step in our proposed scheme is to extract the MCS distribution for the URLLC slice from field measurements. We get this information by implementing an MCS distribution estimation module that collects the MCS from user measurement reports on average during the simulation time and associates them to the gNodeB and slice IDs for constructing a per-gNodeB MCS distribution for the URLLC slice.

Resource dimensioning module

Let $\mathcal{M}_k = \{p_1^{(k)}, ..., p_{|\mathcal{I}|}^{(k)}\}$ be the MCS distribution extracted from the network as mentioned in the previous section; $p_i^{(k)}$ being the probability of having MCS $i \in \mathcal{I}$ in gNodeB k. The resource dimensioning module associates this distribution with the traffic intensity (in packets/msec) to compute the amount of resources to reserve for the URLLC slice to achieve the target QoS. We recall that the QoS is expressed as the percentage of correctly received packets within the delay constraint. We implement the following optimization problem:

$$\min B_u \tag{4.7}$$

subject to the constraint:

$$Pr[D_u(m, B_u) > T_u] \leq \epsilon \tag{4.8}$$

where B_u is, again, the total available resources for URLLC, $D_u(m, B_u)$ is the per-packet delay of equations (4.2, 4.6), T_u is the delay constraint and ϵ is a small positive number.

We solve this stochastic optimization problem using Monte Carlo simulations. In particular, packets arrive at gNodeB k buffer following a Poisson process and have an MCS chosen following the distribution $\mathcal{M}_k$. Their number evolves with time following equation (4.1). The packet's delay is then calculated, leading to the outage probability in (4.8). Note that the packet arrival rate depends on the predicted number of users in the gNodeB.

Once we obtain the packet loss for known B_u, we search for the lowest resource reservation that achieves the packet loss constraint. We show in the following Section 4.3 that the MCS distributions can vary from a gNodeB to another depending on the URLLC users' trajectory, and how this affects the required resource reservation.

4.3 Numerical Results

In our numerical simulations, we simulate the system described in Chapter 3, illustrated in Figure 3.2. We implement our NSSMF proposal described in the previous section and illustrated in Figure 4.1, along with the resource reservation schemes mentioned in Section 3.3.

4.3.1 Model parameters

MCS distributions for two different gNodeBs are shown in Figure 4.2, where we illustrate the probability distribution of the MCS for gNodeB 2 and 10 on one trajectory. We can see that users connected to gNodeB 10 have higher MCS values than those connected to gNodeB 2. It means that users in gNodeB 10 have better radio conditions. The trajectory shown in Figure 3.2 explains this difference, where we can observe that URLLC UEs cross the gNodeB closer to the gNodeB 10 than to gNodeB 2. So we can conclude that gNodeB 2 needs higher RB reservation to compensate for degraded radio conditions.

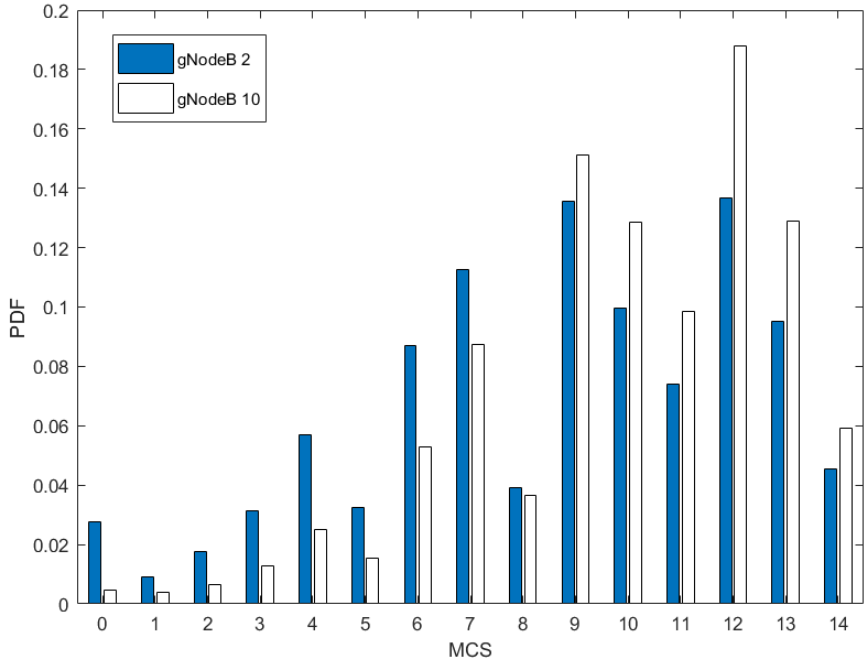

Figure 4.2: MCS distribution.

The resource reservation corresponding to these two gNodeBs is shown in Figure 4.3. We can see the link between the MCS distribution shown in Figure 4.2 and the estimated amount of B_u to be reserved for a certain number of URLLC users. We see that gNodeB 2 should reserve a higher quantity of B_u for a specific number of users than gNodeB 10, which matches its radio conditions.

The vehicular slice SLA requirements are as indicated in the previous

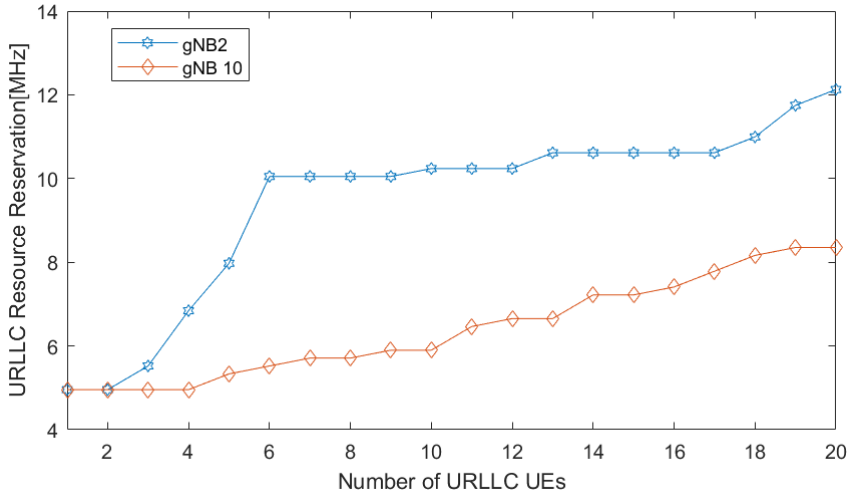

Figure 4.3: Resource reservation per number of users in a gNodeB.

chapter in section 3.2.3. We model the HARQ re-transmission, and we take the following assumptions for latency calculation in equation 4.2 [50]:

- $T_{L1/L2} = 1$ sTTI,
- $T_a = 1$ sTTI,
- $T_{Proc} = 3$ sTTIs
- $T_{tx} = 1$ sTTI.

4.3.2 Performance

In this section, we evaluate the performance of the proactive schemes proposed in Chapter 3 and compare them to the reactive and static ones:

1. *Reactive reservation*, corresponding to the reactive scheme with a reconfiguration delay of 80 ms.
2. *Maximal Static Reservation*, corresponding to a maximal static reservation scheme on all gNodeBs, independent of the number of vehicles in the gNodeB. The quantity of reserved resources ensures, on a worst-case basis, that all vehicles would meet their stringent QoS constraints.
3. *Reservation on neighbors*, corresponding to our first proactive reservation scheme on neighboring gNodeBs.

4. *Trajectory Prediction*, corresponding to our second proactive reservation scheme, making use of predicted trajectory

Figures 4.4 and 4.5 show the average URLLC loss probabilities and the average eMBB throughput, respectively, for the four reservation schemes, for different URLLC users' arrival rates. In our case, these metrics are the most relevant for the considered slices.

We can see that the reactive scheme has a very high URLLC packet loss and the highest eMBB throughput since there is no over-reservation of resources for URLLC users. When an over-reservation of resources is performed in the maximal static scheme, the packet loss of URLLC vehicles reaches very low values at a very high eMBB throughput cost in return. The URLLC packet loss increases when the vehicle arrival rate increases but remains below the target of 10^{-5}. However, the eMBB throughput is independent of the URLLC traffic intensity for the static scheme, as the reservation does not depend on the traffic.

When the reservation is performed on the neighbors and anticipated trajectory, we reach acceptable URLLC packet loss values below 10^{-5}. However, for the eMBB throughput, we can see the negative impact of reserving extra resources for URLLC on neighbors versus the scheme based on the trajectory. The latter enables a high eMBB throughput, almost equivalent to the reactive scheme, and thus achieves the best balance between URLLC reliability and eMBB throughput. With the offline simulation, we see that the resource reservation is not optimal since it counts on a range of user's number in a gNodeB, which leads to under/over resource reservation.

4.4 Conclusion

In this chapter, we continued studying 5G network slicing for vehicular URLLC services under slice reconfiguration delay, as observed in chapter 3. We developed a per-gNodeB slice dimensioning method to assess the required resources to meet vehicle URLLC requirements, based on the knowledge of the gNodeB radio condition distribution and traffic intensity. Our results showed that, with prior knowledge of the trajectory of the vehicular URLLC UEs and the MCS distribution in the different cells, we limited the resource reservation and fulfilled the vehicular URLLC requirements while minimizing the impact on eMBB throughput. These results, when compared to those in chapter 3, give a finer resource reservation, dependent on each gNodeB radio conditions.

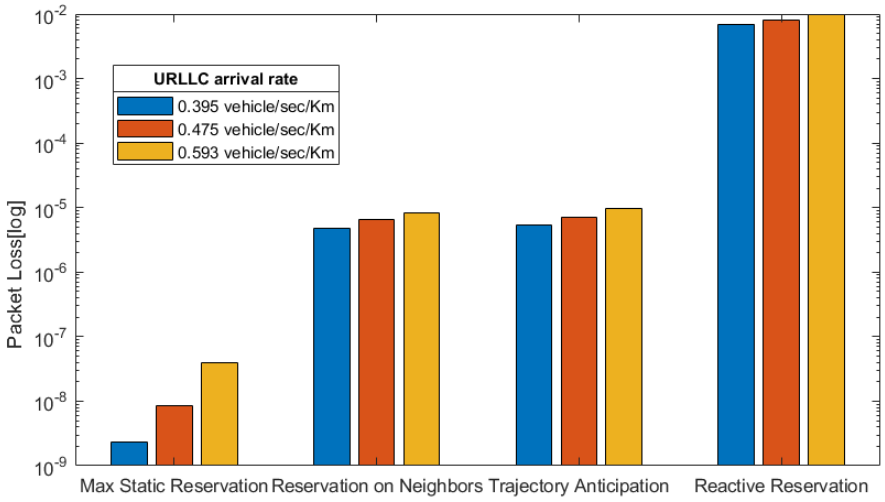

Figure 4.4: URLLC arrival rate impact on reliability for the proposed schemes.

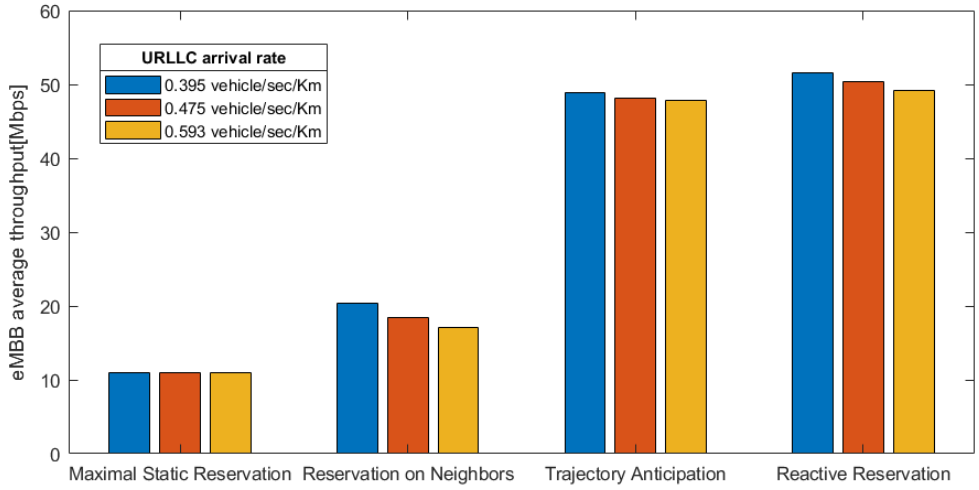

Figure 4.5: URLLC arrival rate impact on eMBB throughput.

Chapter 5

Large deviation bounds for URLLC resource allocation

In the previous chapters, we presented a resource dimensioning method for URLLC and eMBB based on the Markov chain simulation that governs the queue's dynamics. In this chapter, we introduce a mathematical model for computing the outage probability of a URLLC user without relying on time-consuming Markov chain simulations. Our model is based on the large deviations theory [51] and consists in finding an upper bound of the outage probability that corresponds to URLLC maximum latency requirement. Note that this model may be integrated within the resource dimensioning framework of figure 4.1, replacing the simulations in the "Resource dimensioning module" with the upper bound computations presented in this chapter.

5.1 Related works

There have been attempts to use classical queuing theory methods for dimensioning the system, but they needed to make strong hypotheses on the traffic and system. For instance, [52] proposed an M/M/1 model based on the assumption of Poisson arrivals of packets and an exponential model for the variation of packet sizes due to different radio conditions. In [53], the authors use the M/M/m/K queue to model the system reliability for a worst-case scenario where users are assumed to be at the cell edge. An M/G/1 model has been developed in [54], but it concerns the eMBB traffic performance when it is subject to URLLC preemption. [55] relaxed the Exponential assumption for the service rate and adopted an M/G/1 model with vacations but with two restrictive assumptions. First, the "General" service model is due to dif-

ferent packet sizes and not different radio conditions, and second, packets are supposed to be served by one server in continuous time. At the same time, the 5G NR system can multiplex packets in the spectrum dimension (several servers) and is time-slotted. Regarding these limitations and the difficulty in finding realistic and tractable queuing models for URLLC, we adopt a large deviations approach that is suitable to analyze the system's tail, corresponding to the URLLC outage region. We use two types of simulations to validate the model using several bounds suitable for the URLLC sporadic traffic model. We show the tighter bounds in two cases: a very stringent delay budget where the radio procedures do not allow further queuing delays and a less stringent case where several slots are available for queuing within the delay budget. First, we compare the model to numerical simulations of the discrete-time Markov process describing the system evolution, as used in the previous chapters. And next, we show how we implement the dimensioning framework based on the analytical model in the large-scale system-level simulator. Our numerical results show that the derived models can be used for resource dimensioning and do not lead to excessive over-dimensioning.

This chapter is organized as follows: Section 5.2 describes the outage model based on the large deviations theory. Section 5.3 compares the model to numerical resolution based on a radio distribution. Section 5.4 applies the proposed dimensioning framework to the system-level simulator and quantifies the resulting resource reservation gap. Section 5.5 discusses and concludes the work.

5.2 System and traffic model

For developing the analytical model, we consider a 5G gNodeB with U URLLC users and a 5G-NR-like frame, where time/frequency resources are organized into RBs and mini-slots. The slot is of size T ms, and there are R reserved RBs of the total bandwidth dedicated for URLLC traffic. We consider a sporadic traffic model, i.e., a user is active (generates a packet) during a slot with probability q.

As stated before, there are I different MCS, numbered 1 to i, and a packet belongs to a user whose MCS is i with probability p_i. We assume that the MCS distribution in the gNodeB is known. e.g., from field measurements. If a user uses MCS i, each of its packets consumes r_i RBs. Without loss of generality, we suppose that the MCSs are sorted following increasing spectral efficiency, i.e., $r_1 > ... > r_I$.

5.2.1 Outage bounds for a tight delay budget (no waiting)

Let $X_u(t)$ be the number of requested RBs by user $u \in [1, U]$ during slot t. $X_u(t)$ are i.i.d. random variables that take the following values:

$$X_u(t) = \begin{cases} 0, & \text{with prob. } (1-q) \\ r_i & \text{with prob. } qp_i \end{cases} \tag{5.1}$$

The total number of resources requested by packets generated in a given slot is then given by:

$$\bar{R}(t) = \sum_{u=1}^{U} X_u(t) \tag{5.2}$$

The outage occurs when the number of needed resources exceeds the amount of reserved resources. The objective is to ensure that the outage is below a small positive value ϵ:

$$Pr(\sum_{u=1}^{U} X_u > R) \leq \epsilon \tag{5.3}$$

As we can see, problem (5.3) is a large deviation problem for which several bounds exist as a solution. We start by computing the mean and standard deviation of X_u and $\bar{R}$. For X_u, the mean value is:

$$\mu_0 = E[X_u] = q \sum_{i=1}^{I} p_i r_i \tag{5.4}$$

and the variance is:

$$\sigma_0^2 = E[X_u^2] - \mu_0^2 = q \sum_{i=1}^{I} p_i r_i^2 - q^2 (\sum_{i=1}^{I} p_i r_i)^2 \tag{5.5}$$

As for the total consumption of RBs, its mean and variance are $\mu = U\mu_0$ and $\sigma^2 = U\sigma_0^2$, respectively.

Define $x_u = X_u - \mu_0$. The outage constraint (5.3) can be rewritten as:

$$Pr(\sum_{u=1}^{U} x_u > R - \mu) \leq \epsilon \tag{5.6}$$

Define now $s = \frac{R-\mu}{\sigma}$, the constraint can be rewritten as:

$$Pr(\sum_{u=1}^{U} x_u > s\sigma) \leq \epsilon \tag{5.7}$$

Bienaymé-Chebychev bound

The well-known Bienaymé-Chebychev bound [51] can be applied. Taking the bound as equal to ϵ, we have:

$$Pr(\sum_{u=1}^{U} x_u > s\sigma) \leq \frac{1}{s^2}, \tag{5.8}$$

leading to the required reservation:

$$R_1 = \mu + \frac{\sigma}{\sqrt{\epsilon}} \tag{5.9}$$

Bernstein bound

The Bienaymé-Chebychev bound is known to be weak for a sum of random variables. x_i's have the advantage of being independent and bounded, we can apply more tight bounds. Let M be the upper bound of x_i:

$$M = r_0 - q\sum_{i=1}^{I} r_j p_j \tag{5.10}$$

Bernstein [56] proved that the sum of bounded independent random variables is bounded by:

$$Pr(\sum_{u=1}^{U} x_u > s\sigma) \leq \exp\left[-\frac{s^2}{2 + \frac{2}{3}\frac{M}{\sigma}s}\right] \tag{5.11}$$

Substituting the bound by the target, this leads to the reservation:

$$R_2 = \mu - \frac{M\ln\epsilon}{3} + \frac{\sigma}{2}\sqrt{\frac{4M^2(\ln\epsilon)^2}{9\sigma^2} - 8\ln\epsilon} \tag{5.12}$$

Bennet bounds

Bennet [57] proposed two enhancements on Bernstein's bound, as described next.

First, the bound can be computed as:

$$Pr(\sum_{u=1}^{U} x_u > s\sigma) \leq \exp\left[-\frac{s^2}{1 + \frac{1}{3}\frac{M}{\sigma}s + \sqrt{1 + \frac{2}{3}\frac{M}{\sigma}s}}\right] \tag{5.13}$$

Leading to the reservation of resources:

$$R_3 = \sigma s_3 + \mu \tag{5.14}$$

with s_3 solution of the following equation:

$$\frac{s^2}{\ln \epsilon} + 1 + \frac{1}{3}\frac{M}{\sigma}s + \sqrt{1 + \frac{2}{3}\frac{M}{\sigma}s} = 0 \tag{5.15}$$

Bennet [57] also proposed another bound as follows:

$$Pr(\sum_{u=1}^{U} x_u > s\sigma) \leq e^{\frac{s\sigma}{M}} \left(1 + s\frac{M}{\sigma}\right)^{-(\frac{s\sigma}{M} + \frac{\sigma^2}{M^2})} \tag{5.16}$$

Leading to the reservation of resources:

$$R_4 = \sigma s_4 + \mu \tag{5.17}$$

with s_4 solution of the following equation:

$$e^{\frac{s\sigma}{M}} \left(1 + s\frac{M}{\sigma}\right)^{-(\frac{s\sigma}{M} + \frac{\sigma^2}{M^2})} = \epsilon \tag{5.18}$$

5.2.2 Model with queuing

We now consider the case with a looser constraint, as the one studied in the previous chapters, i.e., where a packet can stay for $\delta > 1$ slots in the system before its delay budget expires (e.g., 7 slots for a target delay of 1 ms and a slot length of 0.143 ms). We consider the same traffic model as in the previous section.

Outage probability formulation

In a given slot, numbered 0, knowing that there are R reserved RBs, the "overflow" of resources, i.e. the amount of RBs' that will be needed in the future to serve the backlogged traffic is equal to:

$$B_{(0)} = \left(\sum_{u=1}^{U} X_{(0),u} + B_{(-1)} - R\right)^{+} \tag{5.19}$$

where $X_{(0),u}$ is the amount of resources required for serving the packet of user u generated at slot 0. $B_{(-1)}$ is the amount of overflow traffic from the previous slot (denoted by slot -1), and $(x)^+ = \max(x, 0)$. Recursively, for a previous slot $-j$, the overflow is computed by:

$$B_{(-j)} = \left(\sum_{u=1}^{U} X_{(-j),u} + B_{(-j-1)} - R\right)^{+} \tag{5.20}$$

with $X_{(-j),u}$ being the amount of resources required for serving the packet of user u generated at slot $-j$.

The outage probability is computed by the probability that the new packet has to wait for more than δ slots:

$$Pr(B_{(0)} > \delta R) \leq \epsilon \tag{5.21}$$

Approximate outage probability

We consider a system with memory of m slots, i.e., the probability that packets are waiting from more than m slots is negligible. In this case, we neglect the term $B_{(-m-1)}$ in the overflow. Summing up to the previous m slots, and replacing $\left(\sum_{u=1}^{U} X_{(-j),u} + B_{(-j-1)} - R\right)^+$ by $\sum_{u=1}^{U} X_{(-j),u} + B_{(-j-1)} - R$, the outage constraint becomes:

$$Pr(\sum_{j=1}^{m}\sum_{u=1}^{U} X_{(-j),u} > (\delta + m)R) \leq \epsilon \tag{5.22}$$

This approximation is twofold. First, by neglecting the overflow from slots that are older than m, we suppose that the system is not in overload for a significant time. This assumption is reasonable for the URLLC regime. We will see in the numerical applications that the memory of 10 mini-slots gives a good approximation. Second, by removing the $(.)^+$ operator from the overflow of equation 5.20, we allow the overflow to be negative as if the whole mR resources were used to serve the traffic arriving within the previous m slots. We shall test the validity of this approximation in numerical applications.

The delay constraint (5.22) can then be rewritten by:

$$Pr(\sum_{j=1}^{m}\sum_{u=1}^{U} X_{-(j),u} > (\delta + m + 1)R) \leq \epsilon \tag{5.23}$$

This constraint compares the sum of $U(m+1)$ independent variables with a threshold; it can be rewritten as:

$$Pr(\sum_{j=1}^{m}\sum_{u=1}^{U} x_{-(j),u} > \hat{\sigma}s) \leq \epsilon \tag{5.24}$$

with $\hat{\sigma} = \sqrt{U(m+1)}\sigma_0$ and

$$s = \frac{(\delta + m + 1)R - (m+1)U\mu_0}{\hat{\sigma}}, \tag{5.25}$$

$x_{-(j),u} = X_{-(j),u} - \mu_0$ are centered independent random variables bounded by M computed as in equation (5.10).

We can apply the same bounds of equations (5.8), (5.11), (5.13) and (5.16) on the system.

5.3 Numerical applications

We now compare the analytical bounds with numerical simulations of the scheduler, as used in the previous chapters. We recall here its operation:

- Inputs: the simulator takes as input the traffic profile (number of users, average number of packets per second per user) and the radio conditions. For a realistic setting, we consider in the comparison a typical MCS distribution issued from the system level simulator, as illustrated in Figure 5.1.
- Traffic generation: the time is divided into slots of size $T = 0.143$ ms, and there are R reserved RBs for URLLC. In each slot, each user generates a packet following a Bernoulli law with parameter q, and if a packet is generated, it chooses at random an MCS following the input distribution. Packets are all of equal size (96 bits).
- Scheduler: Packets are served following a FCFS discipline. When a packet is generated, it is put at the end of the queue. A time slot is filled with packets at the head of the queue until all of the R RBs are occupied or the queue is empty. When a packet cannot be scheduled on one slot as the remaining resources are insufficient, it can be scheduled on a consecutive slot.
- Output: Each packet is counted as an outage if the delay between its generation and its service exceeds a threshold.

5.3.1 Model with no waiting

We start with the case of a very stringent delay budget, where there is no room for waiting. We illustrate in Figure 5.2 the outage probability (logarithmic) obtained by simulation, and using the bounds of equations (5.8), (5.11), (5.13) and (5.16). The parameters taken for this simulation are: $U = 20$, $q = 0.072$. First, all the bounds give an outage probability more significant than the simulation. Second, it can be observed that the second bound of Bennet (equation (5.16)) gives the closest bound to the simulation as it is

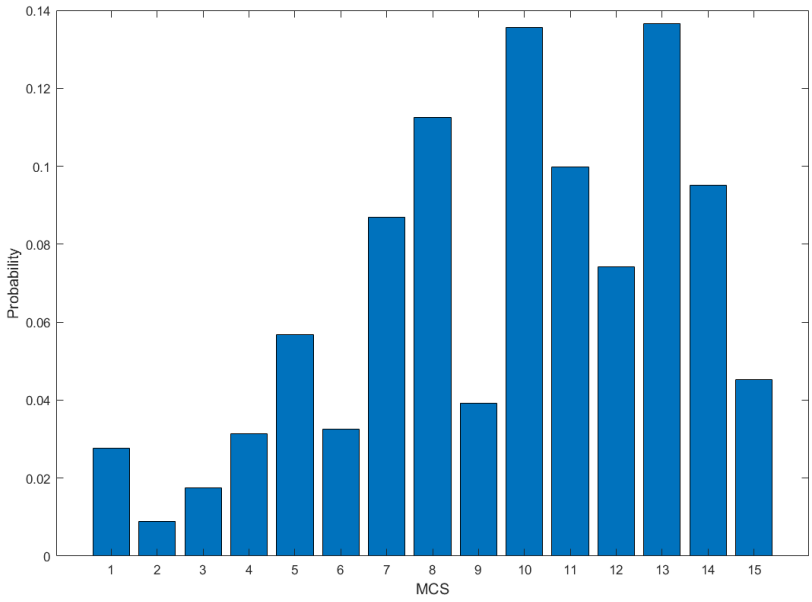

Figure 5.1: MCS distribution.

adapted to a sum of independent variables. Third, the simulation stops for an outage rate below 10^{-7} as the outage event becomes too rare to be simulated.

Based on these results, we investigate the amount of over-dimensioning required when using the analytical bounds compared with the simulation. For a target outage probability of 10^{-5}, the required reservation is of $R = 85$ RBs, based on simulations, while the Bennet bound (5.16) required 115 RBs. The Chebychev bound is so loose that the reservation requirement exceeds 500 RBs.

5.3.2 Model with queuing delay

We now move to a more common use case where there is room for multiple slots for queuing within the delay budget. Here we take the threshold on the waiting delay equal to 1 ms. Note that the threshold depends on the service requirements and the radio settings. The waiting delay threshold has to be computed as the difference between the service delay budget and the other non-compressible delays (alignment, propagation, decoding, back-haul.).

We consider the same MCS distribution as previously. We first start by studying the impact of the approximation of finite memory m on the bound,

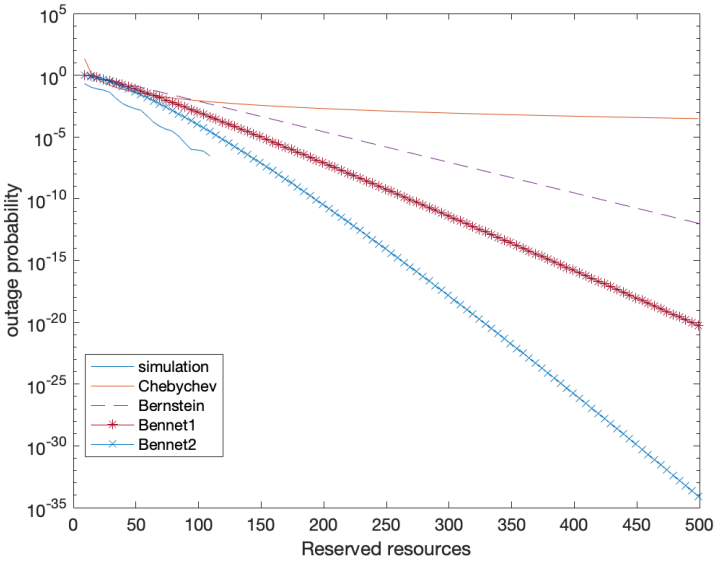

Figure 5.2: Outage probability with no waiting.

considering the tight bound of Bennet (5.16). We can observe in Figure 5.3 that the amount of required reservation increases with m, and stabilizes starting from $m > 9$. We consider in the following $m = 10$.

We compare the analytical bounds with simulation results in Figure 5.4. We see that the difference between the Bennet2 bound achieves the closest bound to the simulation and that the gap is reduced compared to the no-waiting case.

In order to compare with queuing models used in the state of the art, we implement the M/M/c/K model proposed in [53]. The packet arrival is Poisson, service is approximated as exponential, c is the number of servers, and K is the maximum number of packets the system can hold. In [53], they compute K as the number of packets upon arrival that discourages a packet from being queued as it corresponds to an outage ($K = c\delta$ in our case for a fair comparison). However, the number of servers is unknown. They compute the number of packets that can be served in parallel, while this number depends on the MCS for a fixed R. [53] considered the worst case, i.e., when all users are at the gNodeB edge and computed c as the ratio between R and the number of resources occupied by a packet generated at gNodeB edge (MCS 1). As this is too pessimistic, we consider the MCS used by the worst 10% of users (90% percentile), which corresponds to MCS 5. Figure 5.4 shows that the bound is too loose (very large outage). One can

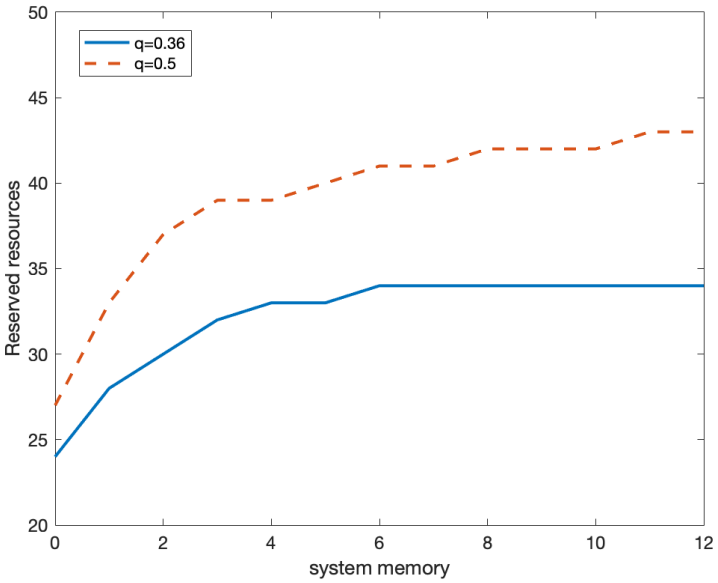

Figure 5.3: Required resource reservation for a target reliability (1 ms budget).

try to consider the average resource consumption instead of the worst case or the percentile ($c = \lceil \frac{R}{\sum_{i=1}^{I} r_i} \rceil$). However, figure 5.4 shows that this method cannot be used for URLLC resource provisioning, as it sometimes largely underestimates the outage (the step-like behavior comes from the necessity to have an integer number of servers in the M/M/c/K model).

The model can also be used for resource dimensioning, i.e., for computing the resource reservation to ensure the target performance. Figure 5.5 compares the amount of reserved resources for the analytical bound (5.16) with the numerical simulations and their outage probability (logarithmic), and shows that the bound is very tight.

5.4 System level simulation

Having validated our analytical model based on simple numerical simulations, we now propose a resource dimensioning framework and test it on the system level simulator.

Figure 4.1 illustrates the architecture for implementing the proposed scheme within the NSSMF. We use the distribution issued from the sys-

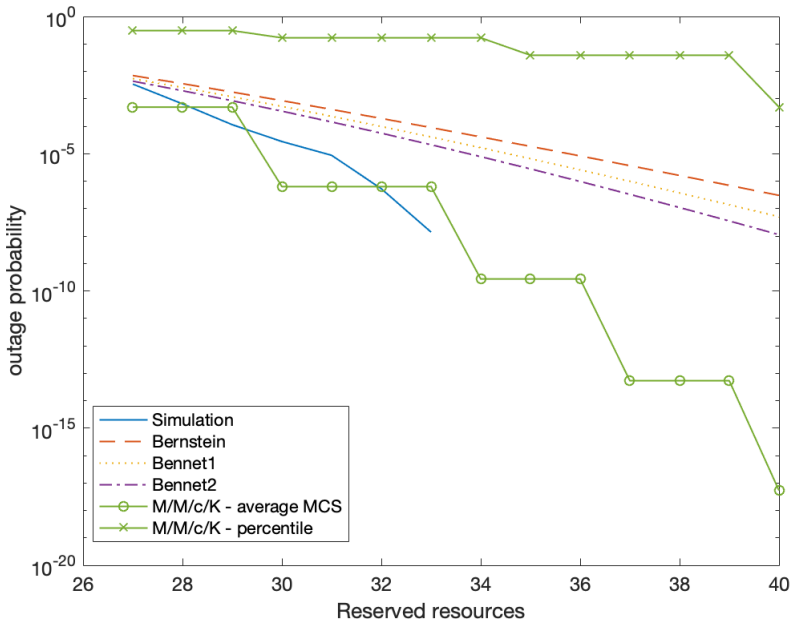

Figure 5.4: Outage probability for the delayed case ($U = 20$, $q = 0.36$).

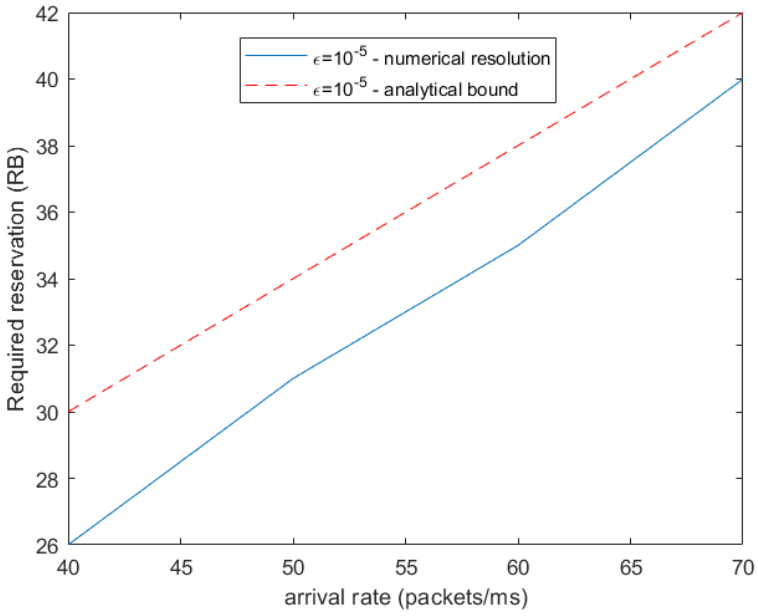

Figure 5.5: Required resource reservation for a target reliability (1 ms budget).

tem level simulator as input for the resource dimensioning module that takes as input the traffic (number of URLLC users, number of packets/user/ms) and computes the needed amount of resources to be reserved for the URLLC slice in each of the gNodeBs, using the analytical model (equation (5.16)). The system applies the new configuration, dynamically changing depending on the NSSMF updates (traffic and radio conditions change).

For the simulations, we implement the system and the dimensioning model described above. The simulated 5G network is in the previous chapter, illustrated in Figure 3.2, where we consider static URLLC users with no eMBB traffic, as our aim is to evaluate the dimensioning method for URLLC.

We perform three types of simulations. The simulation and configuration parameters are presented in Table 5.1 .

Table 5.1: System parameters.

Parameters	URLLC
Environment	3GPP Urban Macro (UMa)
Number of gNodeBs	13
Bandwidth	20 Mhz
SCS	15 Khz
Number of RBs	106
sTTI size(ms)	0.143
Traffic model	Bernoulli
Packet size	96 bits
Speed	Static

In the first simulation, we perform a series of simulations for each traffic intensity, changing the number of reserved resources in each gNodeB until reaching the target of 10^{-5} outage. It gives the system simulation resource reservation, which is not applicable in practice as it requires many trials on the up-and-running network. Second, we apply our dimensioning framework, extracting the radio conditions distribution from the gNodeBs, and then applying the proposed analytical model to obtain the required reservation. Finally, we simulate the M/M/c/K model with 90% percentile MCS. The second set of simulations is based on this analytical reservation (equation (5.16)) to verify that the outage is far below the target. Figure 5.6 compares the reservation obtained by extensive simulations with the analytical model and the M/M/c/K model [53] with a gNodeB edge MCS (worst 10% of users). We first observe that the M/M/c/K model leads to a large over-dimension. As for our proposed bound, we observe an average over-dimensioning ratio of

15% compared to the system simulator, which is acceptable for guaranteeing URLLC reliability, knowing that the bound is computed based only on the knowledge of the average traffic intensity and radio conditions.

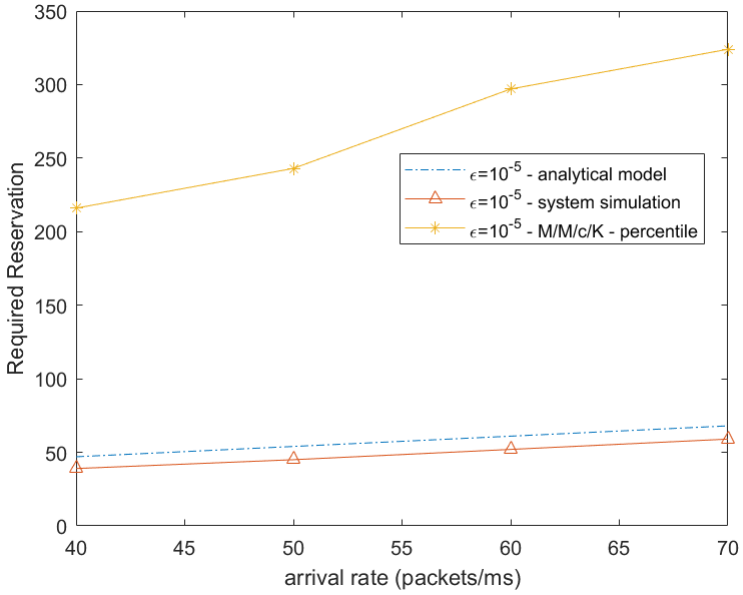

Figure 5.6: System simulations versus analytical model.

5.5 Conclusion and discussions

In this chapter, we extended our dimensioning model presented in chapter 4 with an analytical model. We developed a performance evaluation framework for URLLC traffic in 5G networks based on large deviation bounds. We considered the queuing delay and derived the outage probability bound, i.e., the probability that the delay exceeds a given target.

We first compared the analytical model with the numerical simulation of the scheduler and showed that the proposed bound is tight. We then integrated our model within the resource dimensioning framework presented in the previous chapters. We showed that the URLLC targets are achieved with an acceptable over-dimensioning cost and a low management overhead.

Chapter 6

Conclusion and future perspectives

6.1 Summary

In this thesis, we focused our work on the resource allocation between two primary slices: eMBB and vehicular URLLC. We started with an introduction of the new 5G-NR and its use cases. Afterward, we gave an overview of the mobile network's evolution and the new architecture transformations, detailing the novelty of network slicing and RAN slicing. After that, we summarized the entities responsible for slice creation and management, with a more in-depth description of the radio resource allocation and the new 5G frame structure. Having an optimal resource configuration is the way to guarantee the SLA contract for the mentioned slices. While we might have this configuration at times, with the arrival of vehicular URLLC to a gNodeB, this configuration could be insufficient. So we evoked the problem of reconfiguration of bandwidth and its effect on the URLLC latency since it creates a burst of packet loss due to the time taken by the reconfiguration.

To assess the impact of this reconfiguration, we started by describing the system model and the simulator's block diagram. We simulated the case of reactive resource allocation to see the impact on the packet loss in a gNodeB. So to solve this problem, we proposed two proactive approaches to anticipate the vehicle's arrival in a gNodeB. Depending on the knowledge of the user's trajectory, we apply either the neighboring approach, where we reconfigure all the neighboring gNodeBs and the source gNodeB, or only on the predicted trajectory of the user. We needed an optimal resource

reservation for the URLLC slice to benefit from these approaches. After creating a baseline configuration with offline simulations based on the number of users in a gNodeB, we study another dimensioning model based on queuing model. This dimensioning model inputs the MCS distribution of the different gNodeBs of the network, the latency, and the reliability requirement of the vehicular URLLC slice. Based on the probable number of users per gNodeB and this distribution, packets are generated following Poisson law. To follow, we calculate the number of RBs to be allocated to each packet using two types of schedulers:

- the flexible scheduler that allows the payload to be divided on multiple sTTIs if the remaining resources are insufficient
- the non-flexible scheduler where if the remaining resources are not enough, it will be lost and the packet remains in the queue and is scheduled on the next sTTI

We solved this optimization problem using Monte Carlo simulations to find the optimal reservation that gives us the maximal reliability value. The results of this model were then integrated into the system simulator and used by the scheduler for optimal allocation. Results show that we achieve better resource allocation through finer optimization.

Eventually, we investigated an alternative dimensioning model based on large deviation bounds. We analyzed the tail of the system corresponding to the URLLC outage region. We considered two approaches: with and without packet queuing. We observed that large deviation bounds result in slightly more over-reservation than the system simulations when applied to URLLC, with the advantage of instantaneous computation of the needed resources.

6.2 Perspectives

Slice management at the RAN level is a wide research topic with several axes that can be further investigated in the future. We cite here four research axes that can be studied in future work.

6.2.1 Exploiting geolocation information for URLLC and eMBB resource allocation

In Chapter 4, we exploited the knowledge on the distribution of MCSs in the cell to adjust the resource allocation to the mobile URLLC traffic demand. MCS distributions can be obtained through statistical analysis of historical

data and are supposed to remain the same, on average, over time. This approach presents two limitations: first, it relies on historical data while the traffic can evolve, and second, there could be differences between this average estimation and the exact behavior of the user at a given time. If we suppose that we have complete knowledge of the users' locations and the perceived SINR at each location in real-time, then we can allocate precisely the amount of resources that mobile URLLC users need and maximize the available resources for eMBB users.

Today, geolocation is a research topic that is gaining momentum. As the 5G network is synchronized, the exploitation of metrics related to the arrival time makes it possible to geo-localize users' terminals with high accuracy, 10 m outdoors and down to 3 m indoors as specified by the 3GPP standard [58]. The exploitation of geolocalized measurements to build radio maps, also called radio environment maps (REM), is an old research topic. The REM concept, introduced by [59], consists of spatially interpolating geolocalized measurements to build the whole map of the measured metric. The same concept can be found in the literature under different names, such as Radio Maps [60] or, more recently, as Channel Knowledge Map [61]. Several studies in state of the art focused on building REM, specifically radio coverage maps based on spatial interpolation of geolocalized measurements. Kriging, a spatial interpolation technique widely used in geo statistics, gives the Best Linear Unbiased Prediction (BLUP) at unobserved locations [62]. Kriging provides good accuracy for radio coverage prediction [63–65]. It also provides prediction uncertainty in addition to the average estimates. In [66], the authors extend the coverage prediction with Kriging to predict the SINR and the variance of the error of its estimate based on UE geo-located measurements. The authors show that the predicted SINR distribution (based on the variance of the prediction errors) and its moments are accurate enough that measurement error has no/negligible impact on the average rate prediction of the user when there are enough training measurements.

Hence, by combining SINR maps described above with real-time localization of the users, which will be possible with 5G positioning techniques, we can build an accurate resource reservation for mobile URLLC users and provide maximum resources for eMBB users.

6.2.2 Resource allocation with Artificial Intelligence

ML is one of the Artificial Intelligence (AI) applications that aim to learn and improve some system tasks based on experience and trained data to predict better values without being explicitly programmed. Three machine learning

approaches are introduced depending on the data feedback type: Supervised, Unsupervised, and Reinforcement Learning (RL).

As one of the most known ML techniques, the RL aims to optimize agent decision-making without prior knowledge of the system and environment and labeling input and output data. The agent performs actions based on the environment state, which represents some of its features, and receives feedback regarding the performed actions. This feedback is called a reward or penalty, depending on the agent for taking a good or bad action. The ultimate goal is to maximize the cumulative reward, also referred to as an expected return. Different reinforcement learning methods yield distinct behaviors for the agent to achieve their goal. ML has drawn attention to mobile network research due to its efficiency in addressing optimization problems. This environment and the decision-making process are usually stated as a Markov decision process because many reinforcement learning algorithms use dynamic programming techniques.

However, RL differs because there is no knowledge of the exact mathematical model. Since our model is based on queuing model and gNodeB state, we can apply RL algorithms in future work. We can replace the optimization algorithm presented in chapter 4 with one of the RL algorithms, for example Quality-Learning (Q-Learning) or Deep Reinforcement Learning (DRL). It is one of the most well-known algorithms that maximize the reward in the long term. Many researchers have used DRL to solve resource allocation and optimization problems in mobile networks for eMBB and URLLC slices as in [67], [68], [69] and [70]. We can also mention works studying V2X services for resource allocation in [71], [72] and [73]. To apply the DRL algorithms in our research, we can model our system like the following:

The action taken would be scheduling URLLC packets, and the state of this system can be represented by several elements: the number of waiting packets in a queue, each packet radio conditions, and the packet queuing time. The reward is that the packet waiting time or reliability is less than a certain threshold, or we get a penalty otherwise.

6.2.3 Slice aware traffic steering

Another axis for slice management is traffic steering. It includes intra-layer mobility load balancing and inter-layer traffic steering [74] [75]. In 3GPP standard [76], gNodeB and slice capacity are measured and reported. This information helps design slice-aware mobility load balancing between neighboring gNodeBs and inter-layer mobility.

Those algorithms rely on several parameters and thresholds that must be optimized to adapt to the traffic distribution among slices to offer the best user experience and fulfill the different slice requirements. This optimization can be performed using machine learning techniques. Two approaches can be adopted: predictive techniques (regressions to learn the evolution of the traffic or classification methods to predict congestion) as input for proactive decision-making, or reinforcement learning, where the algorithm learns by interacting with the network.

6.2.4 SLA negotiation

The SLA negotioation is a more global approach where several slice owners have different requirements and are negotiating SLAs with one or several operators [77]. From the operator's point of view, the network should be capable of serving traffic of several coexisting slices while fulfilling the SLA for each of them. The operator should estimate if it can serve a new slice on the existing infrastructure and, if not, evaluate the required upgrades and the related cost.

In [78], a cognitive RAN management framework has been defined for adapting the RAN parameters to the operator objectives using reinforcement learning. This framework can be adapted to determine network strategies that integrate the requirements related to slice-level SLAs. Now, let's look at the SLA negotiation problem. The negotiation process between a set of slice owners and a set of operators can be modeled as a game where each player (operator or vertical) tries to maximize its utility under QoS constraints, or what we call game theory approach for network slicing [79]. This approach helps us meet the operator's requirements by applying constraints on the resources allocated to individual users or users from particular slices.

Bibliography

[1] 3GPP, TS 22.185, *Service requirements for V2X services*, August 2020. version 16.0.0 Release 16.

[2] T. Soni, A. R. Ali, K. Ganesan, and M. Schellmann, "Adaptive numerology—A solution to address the demanding QoS in 5G-V2X," in *2018 IEEE Wireless Communications and Networking Conference (WCNC)*, pp. 1–6, IEEE, 2018.

[3] 5G-PPP, "5g vision." http://5g-ppp.eu/wp-content/uploads/2015/02/5G-Vision-Brochure-v1.pdf, 2015.

[4] M. Bennis, M. Debbah, and H. V. Poor, "Ultra Reliable and Low-Latency Wireless Communication: Tail, Risk, and Scale," *Proceedings of the IEEE*, vol. 106, no. 10, pp. 1834–1853, 2018.

[5] S. K. Sharma and X. Wang, "Toward Massive Machine Type Communications in Ultra-Dense Cellular IoT Networks: Current Issues and Machine Learning-Assisted Solutions," *IEEE Communications Surveys Tutorials*, vol. 22, no. 1, pp. 426–471, 2020.

[6] M. Jiang, M. Condoluci, and T. Mahmoodi, "Network slicing management prioritization in 5g mobile systems," in *European Wireless 2016; 22th European Wireless Conference*, pp. 1–6, 2016.

[7] N. Naddeh, S. Ben Jemaa, S. E. Elayoubi, and T. Chahed, "Proactive RAN Resource Reservation for URLLC Vehicular Slice," in *2021 IEEE 93rd Vehicular Technology Conference (VTC2021-Spring)*, pp. 1–5, 2021.

[8] N. Naddeh, S. Ben Jemaa, S. E. Elayoubi, and T. Chahed, "Anticipatory slice resource reservation for 5g vehicular urllc based on radio statistics," in *2022 IEEE 33rd Annual International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*, pp. 1019–1025, 2022.

[9] 3GPP, TS 23.401, *General Packet Radio Service (GPRS) enhancements for Evolved Universal Terrestrial Radio Access Network (E-UTRAN) access*, 2015. V12.10.0 Release 12.

[10] 3GPP, TS 23.501, *System Architecture for the 5G System*, 2018. V15.1.0 Release 15.

[11] 3GPP, TS 38.300, *NR; NR and NG-RAN Overall description; Stage-2*, 2019. version 15.5.0 Release 15.

[12] S. Yamany and L. M. Contreras, *Network Slicing and Management*, pp. 261–287. Cham: Springer International Publishing, 2021.

[13] T. Wang and S. Wang, "Inter-Slice Radio Resource Allocation: An Online Convex Optimization Approach," *IEEE Wireless Communications*, vol. 28, no. 5, pp. 171–177, 2021.

[14] 3GPP, TR 38.802, *Study on New Radio Access Technology Physical Layer Aspects*, 2017. Version 14.2.0, Release 14.

[15] J. Sachs, G. Wikstrom, T. Dudda, R. Baldemair, and K. Kittichokechai, "5G Radio Network Design for Ultra-Reliable Low-Latency Communication," *IEEE Network*, vol. 32, no. 2, pp. 24–31, 2018.

[16] 3GPP, TS 38.211, *Physical channels and modulation*, July 2020. version 16.2.0 Release 16.

[17] 3GPP, TS 38.101-2, *NR; User Equipment (UE) radio transmission and reception; Part 2: Range 2 Standalone*, 2022. version 17.6.0 Release 17.

[18] H. Chen, R. Abbas, P. Cheng, M. Shirvanimoghaddam, W. Hardjawana, W. Bao, Y. Li, and B. Vucetic, "Ultra-Reliable Low Latency Cellular Networks: Use Cases, Challenges and Approaches," *IEEE Communications Magazine*, vol. 56, no. 12, pp. 119–125, 2018.

[19] NGMN, *5G White Paper*, Feb. 2015. version 1.0.

[20] J. Valgas, D. Martín-Sacristán, and J. Monserrat, "5G New Radio Numerologies and their Impact on V2X Communications," *Waves*, 2018.

[21] 3GPP, TR 38.913, *Study on Scenarios and Requirements for Next Generation Access Technologies*, 2020. version 16.0.0 Release 16.

[22] 3GPP, TS 38.331, *NR; Radio Resource Control (RRC) protocol specification*, 2018. version 15.3.0 Release 15.

[23] K. Boutiba, A. Ksentini, B. Brik, Y. Challal, and A. Balla, "NRflex: Enforcing network slicing in 5G New Radio," *Computer Communications*, vol. 181, 10 2021.

[24] X. Lin, D. Yu, and H. Wiemann, "A Primer on Bandwidth Parts in 5G New Radio," April 2020.

[25] J. Pérez-Romero, O. Sallent, R. Ferrús, and R. Agustí, "On the configuration of radio resource management in a sliced RAN," in *NOMS 2018 - 2018 IEEE/IFIP Network Operations and Management Symposium*, pp. 1–6, 2018.

[26] L. Feng, Y. Zi, W. Li, F. Zhou, P. Yu, and M. Kadoch, "Dynamic Resource Allocation With RAN Slicing and Scheduling for uRLLC and eMBB Hybrid Services," *IEEE Access*, vol. 8, pp. 34538–34551, 2020.

[27] J. Kwak, J. Moon, H.-W. Lee, and L. B. Le, "Dynamic network slicing and resource allocation for heterogeneous wireless services," in *2017 IEEE 28th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC)*, pp. 1–5, 2017.

[28] T. Mumtaz, S. Muhammad, M. I. Aslam, and I. Ahmed, "Inter-slice resource management for 5G radio access network using markov decision process," *Telecommunication Systems*, vol. 79, pp. 541–557, Apr 2022.

[29] B. Khodapanah, A. Awada, I. Viering, J. Francis, M. Simsek, and G. P. Fettweis, "Radio Resource Management in context of Network Slicing: What is Missing in Existing Mechanisms?," in *2019 IEEE Wireless Communications and Networking Conference (WCNC)*, pp. 1–7, 2019.

[30] P. Korrai, E. Lagunas, S. K. Sharma, S. Chatzinotas, A. Bandi, and B. Ottersten, "A RAN Resource Slicing Mechanism for Multiplexing of eMBB and URLLC Services in OFDMA based 5G Wireless Networks," *IEEE Access*, vol. 8, pp. 45674–45688, 2020.

[31] A. Azari, M. Ozger, and C. Cavdar, "Risk-Aware Resource Allocation for URLLC: Challenges and Strategies with Machine Learning," *IEEE Communications Magazine*, vol. 57, 12 2018.

[32] M. Al-Ali, E. Yaacoub, and A. Mohamed, "Dynamic resource allocation of embb-urllc traffic in 5g new radio," in *2020 IEEE International Conference on Advanced Networks and Telecommunications Systems (ANTS)*, pp. 1–6, 2020.

[33] A. Destounis, G. S. Paschos, J. Arnau, and M. Kountouris, "Scheduling URLLC users with reliable latency guarantees," in *2018 16th Interna-*

tional Symposium on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks (WiOpt), pp. 1–8, 2018.

[34] E. Fountoulakis, N. Pappas, Q. Liao, V. Suryaprakash, and D. Yuan, "An examination of the benefits of scalable TTI for heterogeneous traffic management in 5G networks," in *2017 15th International Symposium on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks (WiOpt)*, pp. 1–6, 2017.

[35] K. I. Pedersen, G. Berardinelli, F. Frederiksen, P. Mogensen, and A. Szufarska, "A flexible 5G frame structure design for frequency-division duplex cases," *IEEE Communications Magazine*, vol. 54, no. 3, pp. 53–59, 2016.

[36] Z. Li, M. A. Uusitalo, H. Shariatmadari, and B. Singh, "5G URLLC: Design challenges and system concepts," in *2018 15th international symposium on wireless communication systems (ISWCS)*, pp. 1–6, IEEE, 2018.

[37] Y. Han, S. E. Elayoubi, A. Galindo-Serrano, V. S. Varma, and M. Messai, "Periodic radio resource allocation to meet latency and reliability requirements in 5G networks," in *2018 IEEE 87th Vehicular Technology Conference (VTC Spring)*, pp. 1–6, IEEE, 2018.

[38] Q. He, G. Dán, and G. P. Koudouridis, "Semi-Persistent Scheduling for 5G Downlink based on Short-Term Traffic Prediction," in *GLOBECOM 2020 - 2020 IEEE Global Communications Conference*, pp. 1–6, 2020.

[39] H. D. R. Albonda and J. Pérez-Romero, "An Efficient RAN Slicing Strategy for a Heterogeneous Network with eMBB and V2X services," *IEEE Access*, vol. 7, pp. 44771–44782, 2019.

[40] B. Singh, O. Tirkkonen, Z. Li, and M. A. Uusitalo, "Contention-Based Access for Ultra-Reliable Low Latency Uplink Transmissions," *IEEE Wireless Communications Letters*, vol. 7, pp. 182–185, April 2018.

[41] S. E. Elayoubi, P. Brown, M. Deghel, and A. Galindo-Serrano, "Radio Resource Allocation and Retransmission Schemes for URLLC over 5G networks," *IEEE JSAC*, vol. 37, pp. 896–904, April 2019.

[42] K. I. Pedersen, G. Pocovi, J. Steiner, and S. R. Khosravirad, "Punctured Scheduling for Critical Low Latency Data on a Shared Channel with Mobile Broadband," in *2017 IEEE 86th Vehicular Technology Conference (VTC-Fall)*, pp. 1–6, 2017.

[43] A. Anand, G. De Veciana, and S. Shakkottai, "Joint Scheduling of URLLC and eMBB Traffic in 5G Wireless Networks," in *IEEE INFOCOM 2018 - IEEE Conference on Computer Communications*, pp. 1970–1978, 2018.

[44] M. Morcos, M. Mhedhbi, A. Galindo-Serrano, and S. Eddine Elayoubi, "Optimal resource preemption for aperiodic URLLC traffic in 5G Networks," in *2020 IEEE 31st Annual International Symposium on Personal, Indoor and Mobile Radio Communications*, pp. 1–6, 2020.

[45] Y. Chen, L. Cheng, and L. Wang, "Prioritized resource reservation for reducing random access delay in 5G URLLC," in *2017 IEEE 28th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC)*, pp. 1–5, 2017.

[46] S. E. Elayoubi, S. Ben Jemaa, Z. Altman, and A. Galindo-Serrano, "5G RAN slicing for verticals: Enablers and challenges," *IEEE Communications Magazine*, vol. 57, no. 1, pp. 28–34, 2019.

[47] A. Chagdali, S. E. Elayoubi, and A. M. Masucci, "Slice Function Placement Impact on the Performance of URLLC with Multi-Connectivity," *Computers*, vol. 10, no. 5, p. 67, 2021.

[48] 3GPP, TR 36.814, *Evolved Universal Terrestrial Radio Access (E-UTRA); Further advancements for E-UTRA physical layer aspects*, Jan. 2010. version 1.6.0 Release 9.

[49] 3GPP, TS 38.101-1, *Study on NR Vehicle-to-Everything (V2X)*, 2020. version 16.4.0 Release 16.

[50] 3GPP, TDoc R1-1802882, *Latency for URLLC*, Feb. 2018. version 16.0.0 Release 16.

[51] D. W. Stroock, *An introduction to the theory of large deviations*. Springer Science & Business Media, 2012.

[52] A. Chagdali, S. E. Elayoubi, A. M. Masucci, and A. Simonian, "Performance of URLLC Traffic Scheduling Policies with Redundancy," in *2020 32nd International Teletraffic Congress (ITC 32)*, pp. 55–63, IEEE, 2020.

[53] C.-P. Li, J. Jiang, W. Chen, T. Ji, and J. Smee, "5g ultra-reliable and low-latency systems design," in *2017 European Conference on Networks and Communications (EuCNC)*, pp. 1–5, 2017.

[54] Y. Kim and S. Park, "Calculation method of spectrum requirement for IMT-2020 eMBB and URLLC with puncturing based on M/G/1 priority queuing model," *IEEE Access*, vol. 8, pp. 25027–25040, 2020.

[55] H. Jang, J. Kim, W. Yoo, and J.-M. Chung, "URLLC mode optimal resource allocation to support HARQ in 5G wireless networks," *IEEE Access*, vol. 8, pp. 126797–126804, 2020.

[56] S. Bernštein, "Theory of probability," *Moscow. MR0169758*, 1927.

[57] G. Bennett, "Probability inequalities for the sum of independent random variables," *Journal of the American Statistical Association*, vol. 57, no. 297, pp. 33–45, 1962.

[58] 3GPP, TR 38.855, *Study on NR positioning support*, 2019. Version 16.0.0, Release 16.

[59] Y. Zhao, J. H. Reed, S. Mao, and K. K. Bae, "Overhead analysis for radio environment map enabled cognitive radio networks," in *2006 1st IEEE Workshop on Networking Technologies for Software Defined Radio Networks*, pp. 18–25, IEEE, 2006.

[60] J. B. et. al., "Fast Radio Map Construction by using Adaptive Path Loss Model Interpolation in Large-Scale Building," *Sensors*, 2019.

[61] Y. Zeng and X. Xu, "Toward Environment-Aware 6G Communications via Channel Knowledge Map," *IEEE Wireless Communications*, 2021.

[62] N. Cressies, "Statistics for spatial data," *Terra Nova*, 1992.

[63] J. Riihijarvi, P. Mahonen, M. Wellens, *et al.*, "Characterization and modelling of spectrum for dynamic spectrum access with spatial statistics and random fields," in *IEEE 19th International Symposium on Personal, Indoor and Mobile Radio Communications,*, pp. 1–6, 2008.

[64] H. Braham, S. Ben Jemaa, G. Fort, *et al.*, "Fixed Rank Kriging for Cellular Coverage Analysis," *IEEE Transactions on Vehicular Technology*, vol. 66, no. 5, pp. 4212–4222, 2017.

[65] A. M. Alam, S. Benjemaa, and T. Romary, "Performance Evaluation of Covariance Tapering for Coverage Mapping," *87th IEEE Vehicular Technology Conference (VTC spring)*, 2018.

[66] I. Hadj Kacem, S. Benjemaa, H. Braham, and A. M. Alam, "SINR Prediction in Presence of Correlated Shadowing in Cellular Networks," *IEEE Transaction on Wireless Communications*, vol. 21, no. 10, 2022.

[67] M. Alsenwi, N. Tran, M. Bennis, S. Pandey, A. Bairagi, and C. S. Hong, "Intelligent Resource Slicing for eMBB and URLLC Coexistence in 5G and Beyond: A Deep Reinforcement Learning Based Approach," *IEEE Transactions on Wireless Communications*, vol. PP, pp. 1–1, 02 2021.

[68] Y. Huang, S. Li, C. Li, Y. Hou, and W. Lou, "A Deep Reinforcement Learning-based Approach to Dynamic eMBB/URLLC Multiplexing in 5G NR," *IEEE Internet of Things Journal*, vol. PP, pp. 1–1, 03 2020.

[69] Y.-H. Hsu and W. Liao, "eMBB and URLLC Service Multiplexing Based on Deep Reinforcement Learning in 5G and Beyond," in *2022 IEEE Wireless Communications and Networking Conference (WCNC)*, pp. 1467–1472, 2022.

[70] J. Tian, Q. Liu, H. Zhang, and D. Wu, "Multiagent Deep-Reinforcement-Learning-Based Resource Allocation for Heterogeneous QoS Guarantees for Vehicular Networks," *IEEE Internet of Things Journal*, vol. 9, no. 3, pp. 1683–1695, 2022.

[71] J. Li, J. Zhao, and X. Sun, "Deep Reinforcement Learning Based Wireless Resource Allocation for V2X Communications," in *2021 13th International Conference on Wireless Communications and Signal Processing (WCSP)*, pp. 1–5, 2021.

[72] M. Chen, J. Chen, X. Chen, S. Zhang, and S. Xu, "A Deep Learning Based Resource Allocation Scheme in Vehicular Communication Systems," in *2019 IEEE Wireless Communications and Networking Conference (WCNC)*, pp. 1–6, 2019.

[73] X. Zhang, M. Peng, S. Yan, and Y. Sun, "Deep-Reinforcement-Learning-Based Mode Selection and Resource Allocation for Cellular V2X Communications," *IEEE Internet of Things Journal*, vol. 7, no. 7, pp. 6380–6391, 2020.

[74] F. Kavehmadavani, V.-D. Nguyen, T. X. Vu, and S. Chatzinotas, "Traffic Steering for eMBB and uRLLC Coexistence in Open Radio Access Networks," in *2022 IEEE International Conference on Communications Workshops (ICC Workshops)*, pp. 242–247, 2022.

[75] A. Chagdali, S. E. Elayoubi, and A. M. Masucci, "Impact of Slice Function Placement on the Performance of URLLC with Redundant Coverage," in *2020 16th International Conference on Wireless and Mobile Computing, Networking and Communications (WiMob)*, pp. 1–6, 2020.

[76] 3GPP, *NG-RAN; Xn Application Protocol (XnAP)*, 4 2021. version 16.5.0 Release 16.

[77] A. Al Falasi, M. Serhani, and Y. Hamdouch, "A Game Theory Based Automated SLA Negotiation Model for Confined Federated Clouds," 10 2015.

[78] T. Daher, S. Ben Jemaa, and L. Decreusefond, "Cognitive management of self — organized radio networks based on multi armed bandit," in *2017 IEEE 28th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC)*, pp. 1–5, 2017.

[79] X. Yang, Y. Liu, K. S. Chou, and L. Cuthbert, "A game-theoretic approach to network slicing," in *2017 27th International Telecommunication Networks and Applications Conference (ITNAC)*, pp. 1–4, 2017.

ECOLE
DOCTORALE

Titre : Impact du slicing sur la gestion des ressources radio en 5G pour les URLLC véhiculaires et eMBB

Mots clés : 5G, RRM, Slicing, Allocation de ressources, dimensionnement, service vehiculaire

Résumé : La 5G-NR (Fifth Generation-New Radio) a introduit le concept de slicing pour cibler différents types de services. Nous considérons dans cette thèse le trafic véhiculaire, les véhicules envoyant deux types de flux : eMBB (enhanced Mobile BroadBand) et URLLC (Ultra-Reliable and Low Latency Communications). Ces flux sont acheminés en deux slices différents, la première cherchant à garantir et/ou maximiser le débit, tandis que la seconde doit répondre à de fortes contraintes de QoS(Quality of Service) en termes de délai, de l'ordre de 1ms, et de fiabilité, sur de l'ordre de 99.999%. Ces slices avec des profils de trafic et des exigences de QoS hétérogènes doivent partager la même infrastructure physique. Cette thèse vise à proposer de nouveaux schémas d'allocation de ressources pour satisfaire les exigences strictes de qualité de service de l'URLLC sans impacter trop le trafic eMBB. L'un des principaux défis est le moment où les ressources initialement réservées à l'eMBB doivent être allouées à l'arrivée de nouveaux flux URLLC. En raison de l'utilisation de différentes numérologies, ces ressources doivent être reconfigurées, ce qui ajoute un délai supplémentaire de l'ordre de 80 ms, ce qui dépasse le budget de délai URLLC. Pour répondre à ce problème de délai, nous proposons des schémas proactifs de réservation de ressources pour URLLC qui anticipent l'arrivée des véhicules dans une cellule et (re-)configurent la tranche avant leur arrivée effective dans la cellule. Ces approches permettent de répondre aux exigences de délai et de débit du trafic URLLC et eMBB des véhicules, respectivement. Nous introduisons en outre un modèle de dimensionnement inter-slice qui prend en compte les conditions radio et les trajectoires de l'utilisateur dans le réseau, ce qui permet de prendre en compte les MCS (Modulation and Coding Scheme) des utilisateurs. Ce faisant, nous obtenons une meilleure allocation des ressources grâce à une optimisation plus fine. Nos résultats montrent que nous sommes en mesure de satisfaire les exigences de trafic avec une meilleure utilisation des ressources. Finalement, nous étudions un modèle de dimensionnement alternatif basé sur des bornes de grande déviation. Nous analysons la queue du système correspondant à la région de perte URLLC. Nous considérons deux approches : avec et sans mise en file d'attente de paquets. Nous observons que les grandes limites d'écart entraînent une surréservation légèrement supérieure à l'approche susmentionnée lorsqu'elle est appliquée à l'URLLC, avec l'avantage du calcul instantané des ressources nécessaires.

Title : Impact of slicing on radio resource management in 5G for vehicular URLLC and eMBB

Keywords : 5G, RRM, Slicing, Resource allocation , dimensioning, vehicular service

Abstract :
The Fifth Generation-New Radio (5G-NR) introduced the concept of slicing to target different types of services. We consider in this thesis vehicular traffic, with vehicles sending two types of flows : enhanced Mobile BroadBand (eMBB) and Ultra-Reliable and Low Latency Communications (URLLC). These flows are transported in two different slices, the former trying to guarantee and/or maximize the throughput, while the latter has to meet stringent Quality of Service (QoS) constraints in terms of delay, on the order of 1ms, and reliability, on the order of 99,999%. These slices with heterogeneous traffic profiles and QoS requirements must share the same physical infrastructure. This thesis aims to propose new resource allocation schemes to satisfy URLLC stringent QoS requirements without impacting too much eMBB traffic. One main challenge is when resources initially reserved for eMBB must be allocated to the arrival of new URLLC flow. Due to using different numerologies, these resources need to be reconfigured, adding extra delay on the order of 80ms, which exceeds the URLLC delay budget. To respond to this delay problem, we propose proactive resource reservation schemes for URLLC which anticipates the vehicles' arrival in a cell and (re-)configures the slice before their effective arrival in the cell. These approaches enable to meet the delay and throughput requirements of vehicular URLLC and eMBB traffic, respectively. We additionally introduce an inter-slice dimensioning model that considers user's radio conditions and trajectories in the network, which enables taking into consideration users Modulation and Coding Schemes (MCS). By doing so, we achieve a better resource allocation through finer optimization. Our results show that we are able to satisfy traffic requirements with a better resource utilization. Eventually, we investigate an alternative dimensioning model based on large deviation bounds. We analyze the tail of the system corresponding to the URLLC outage region. We consider two approaches : with and without packet queuing. We observe that large deviation bounds result in slightly more over-reservation than the aforementioned approach when applied to URLLC, with the advantage of instantaneous computation of the needed resources.

Institut Polytechnique de Paris
91120 Palaiseau, France