Latency reduction in narrowband cellular networks : applications to IoT and V2X

Zubair Amjad

Latency Reduction in Narrowband Cellular Networks: Applications to IoT and V2X

A thesis for the academic degree of
Doctor of Philosophy
submitted by
M.Eng.
Zubair AMJAD

IRIMAS, Université de Haute-Alsace, Mulhouse/Colmar
ivESK, Offenburg University of Applied Sciences

Evaluation committee:

Head of the Jury: Pr. Bertrand DUCOURTHIAL, Universitè de Technologie de Compiègne

Reviewer: Dr. Hdr Oumaya BAALA, University of Technology of Belfort-Montbéliard

Reviewer: Pr. Dr. Christian SCHINDELHAUER, Albert-Ludwigs-Universität Freiburg

Thesis director: Dr. Hdr Benoît HILT, Université de Haute-Alsace

Co-supervisor: Pr. Dr.-Ing. Axel SIKORA, Hochschule Offenburg

Co-supervisor: Pr. Jean-Philippe LAUFFENBURGER, Université de Haute-Alsace

Date of dissertation presentation: 26^{th} May, 2020

Mulhouse 2020

Résumé

L'évolution des réseaux cellulaires de la première génération (1G) à la quatrième génération (4G) a été dictée par la demande de capacité de liaison descendante centrée sur l'utilisateur, également appelée techniquement large bande mobile"(MBB – Mobile Broadband). Avec sa cinquième génération (5G), la communication de type machine (MTC – Machine Type Communication) a été ajoutée dans les cas d'utilisation cibles et la prochaine génération de réseaux cellulaires devrait les prendre en charge. Mais cette dernière prise en charge nécessite des améliorations des technologies existantes en termes de latence, de fiabilité, d'efficacité énergétique, de débit de données, d'extensibilité et de capacité.

À l'origine, le MTC a été conçu pour des applications à faible bande passante et à forte latence, telles que la détection environnementale, la poubelle intelligente, etc. Aujourd'hui, la demande s'est accrue pour les applications nécessitant une faible latence. Parmi les autres défis bien connus des réseaux cellulaires récents, tels que l'efficacité énergétique du débit de données, la fiabilité, etc., la latence n'est pas non plus adaptée aux applications critiques telles que le contrôle en temps réel des machines, la conduite autonome, l'Internet tactile, etc. Par conséquent, dans les réseaux cellulaires actuellement déployés, il est nécessaire de réduire la latence et d'augmenter la fiabilité offerte par les réseaux pour prendre en charge des cas d'utilisation tels que la conduite autonome coopérative ou l'automatisation industrielle, qui sont regroupés sous la dénomination de communication ultra fiable à faible latence (URLLC – Ultra Reliable Low Latency Communication).

Cette thèse porte principalement sur la latence dans le réseau d'accès radio terrestre universel (UTRAN – Universal Terrestrial Radio Access Network) des réseaux cellulaires. L'ensemble du travail est divisé en cinq parties. La première partie présente l'état de l'art pour les réseaux cellulaires. La deuxième partie contient un aperçu détaillé des cas d'utilisation des URLLC et des exigences que doivent remplir les réseaux cellulaires pour les prendre en charge. Le travail de cette thèse est réalisé dans le cadre d'un projet de collaboration entre le laboratoire IRIMAS de l'Université de Haute-Alsace, France et l'Institut pour des systèmes embarqués et une électronique de communication fiables (ivESK) de l'Université des sciences appliquées d'Offenburg, Allemagne. Les cas d'utilisation sélectionnés de l'URLLC font partie des intérêts de recherche des deux instituts partenaires. La troisième partie présente une étude et une évaluation détaillées des mécanismes de latence des plans de contrôle et des utilisateurs dans la génération actuelle des réseaux cellulaires. L'évaluation et l'analyse de ces latences, réalisées avec le simulateur ns-3 open-source, ont été menées en explorant un large éventail de paramètres qui comprennent entre autres, des modèles de trafic, des paramètres d'accès aux canaux, des modèles de propagation réalistes, et un large ensemble de paramètres de

pile de protocoles de réseaux cellulaires. Ces simulations ont été réalisées avec des appareils à faible puissance, à faible coût et à large portée, communément appelés appareils IoT, et normalisés pour les réseaux cellulaires. Ces dispositifs utilisent les technologies LTE-M ou Narrowband-IoT (NB-IoT) qui sont conçues pour les choses connectées. Ils se distinguent principalement par la largeur de bande fournie et d'autres caractéristiques supplémentaires telles que le schéma de codage, la complexité du dispositif, etc.

La quatrième partie de cette thèse présente une étude, une mise en œuvre et une évaluation des techniques de réduction de la latence qui ciblent les différentes couches de la pile de protocoles du réseau LTE (Long Term Evolution) actuellement utilisée. Ces techniques basées sur la réduction de l'intervalle de temps de transmission (TTI – Transmission Time Interval) et les méthodes d'ordonnancement semi-persistant (SPS – Semi-Persistent Scheduling) sont implémentées dans le simulateur ns-3 et sont évaluées par des simulations réalistes réalisées pour une variété de cas d'utilisation à faible latence axés sur l'automatisation industrielle et la mise en réseau de véhicules. Pour tester les techniques de réduction de latence proposées dans les réseaux cellulaires, étant donné que ns-3 ne prend pas en charge le NB-IoT dans sa version actuelle, une extension du NB-IoT pour le module LTE a été développée. Cela permet d'explorer les limites et les problèmes de déploiement.

Dans la dernière partie de cette thèse, un cadre de déploiement flexible appelé Hybrid Scheduling and Flexible TTI pour les techniques de réduction de latence proposées est présenté, mis en œuvre et évalué par des simulations réalistes. À l'aide de l'évaluation de la simulation, il est montré que le réseau LTE amélioré proposé et mis en œuvre dans le simulateur peut prendre en charge des applications à faible latence avec des dispositifs à faible coût, à plus grande portée et à bande passante étroite. Les travaux de cette thèse mettent en évidence les techniques d'amélioration potentielles, leurs problèmes de déploiement et ouvrent la voie au support des applications URLLC avec les réseaux cellulaires à venir.

Abstract

The evolution of cellular networks from its first generation (1G) to its fourth generation (4G) was driven by the demand of user-centric downlink capacity also technically called Mobile Broad-Band (MBB). With its fifth generation (5G), Machine Type Communication (MTC) has been added into the target use cases and the upcoming generation of cellular networks is expected to support them. However, such support requires improvements in the existing technologies in terms of latency, reliability, energy efficiency, data rate, scalability, and capacity.

Originally, MTC was designed for low-bandwidth high-latency applications such as, environmental sensing, smart dustbin, etc. Nowadays there is an additional demand around applications with low-latency requirements. Among other well-known challenges for recent cellular networks such as data rate energy efficiency, reliability etc., latency is also not suitable for mission-critical applications such as real-time control of machines, autonomous driving, tactile Internet etc. Therefore, in the currently deployed cellular networks, there is a necessity to reduce the latency and increase the reliability offered by the networks to support use cases such as, cooperative autonomous driving or factory automation, that are grouped under the denomination Ultra-Reliable Low-Latency Communication (URLLC).

This thesis is primarily concerned with the latency into the Universal Terrestrial Radio Access Network (UTRAN) of cellular networks. The overall work is divided into five parts. The first part presents the state of the art for cellular networks. The second part contains a detailed overview of URLLC use cases and the requirements that must be fulfilled by the cellular networks to support them. The work in this thesis is done as part of a collaboration project between IRIMAS lab in Université de Haute-Alsace, France and Institute for Reliable Embedded Systems and Communication Electronics (ivESK) in Offenburg University of Applied Sciences, Germany. The selected use cases of URLLC are part of the research interests

of both partner institutes. The third part presents a detailed study and evaluation of user- and control-plane latency mechanisms in current generation of cellular networks. The evaluation and analysis of these latencies, performed with the open-source ns-3 simulator, were conducted by exploring a broad range of parameters that include among others, traffic models, channel access parameters, realistic propagation models, and a broad set of cellular network protocol stack parameters. These simulations were performed with low-power, low-cost, and wide-range devices, commonly called IoT devices, and standardized for cellular networks. These devices use either LTE-M or Narrowband-IoT (NB-IoT) technologies that are designed for connected things. They differ mainly by the provided bandwidth and other additional characteristics such as coding scheme, device complexity, and so on.

The fourth part of this thesis shows a study, an implementation, and an evaluation of latency reduction techniques that target the different layers of the currently used Long Term Evolution (LTE) network protocol stack. These techniques based on Transmission Time Interval (TTI) reduction and Semi-Persistent Scheduling (SPS) methods are implemented into the ns-3 simulator and are evaluated through realistic simulations performed for a variety of low-latency use cases focused on industry automation and vehicular networking. For testing the proposed latency reduction techniques in cellular networks, since ns-3 does not support NB-IoT in its current release, an NB-IoT extension for LTE module was developed. This makes it possible to explore deployment limitations and issues.

In the last part of this thesis, a flexible deployment framework called Hybrid Scheduling and Flexible TTI for the proposed latency reduction techniques is presented, implemented and evaluated through realistic simulations. With help of the simulation evaluation, it is shown that the improved LTE network proposed and implemented in the simulator can support low-latency applications with low cost, higher range, and narrow bandwidth devices. The work in this thesis points out the potential improvement techniques, their deployment issues and paves the way towards the support for URLLC applications with upcoming cellular networks.

Dedicated to

my dearest parents, loving brother, beloved wife, and lovely son.

Acknowledgements

Foremost, I would like to present my humble gratitude to God Almighty who bestowed me with the courage and ability to come this far and make accomplishments in the noble cause of learning and serving. It is a great pleasure to extend my heartiest thanks to all those who have provided guidance and extended a helping hand to me and have contributed directly or indirectly in this work.

I feel greatly indebted to my thesis director *Dr. Benoit Hilt*, and co-supervisors *Prof. Dr.-Ing. Axel Sikora* and *Prof. Jean-Philippe Lauffenburger* for giving me the opportunity to pursue a doctoral degree and for their scientific supervision throughout the research work. This work would not have been possible without their guidance, encouragement and keen interest. I also wish to extend my gratitude to *Dr. Oumaya Baala* and *Prof. Dr. Cristian Schindelhauer* for their helpful comments and suggestions during thesis review. I would also like to thank *Prof. Dr.-Ing. Tobias Felhauer* and *Prof. Bertrand Ducourthial* for reviewing my work during second year of Ph.D. and serving as the president of the thesis defense jury respectively.

Since this thesis was offered as a collaboration between two research institutes, ivESK at Offenburg University of Applied Sciences (HSO) and IRIMAS at Université de Haute-Alsace (UHA), I had the opportunity to worked with colleagues from both teams. This has been an excellent experience in terms of learning and self-development. I would like to thank all the colleagues from both institutes. All of them provided a creative and inspiring working atmosphere during the past years and contributed in one or the other way to this work. I would specially like to thank *Manuel Schappacher* and *Andreas Walz* from ivESK, for their guidance not only in my work but also for the administrative steps at HSO. I thank *Jonathan Ledy* from IRIMAS for his extended support in administrative steps during my initial few months at UHA. I would also like to extend my gratitude to *Julia Junker* for her continuous behind the scenes support at ivESK. I had interesting discussions and ideas from colleagues at Hahn-Schickard, Commsolid GmbH, and Fraunhofer HHI while working on Taktilus project.

I am thankful to my many colleagues/friends for being part of this journey. I am especially grateful to *Dr. Afaq Muhammad* and *Suleman Ayub* for their enormous help and guidance during these years. I also thank my colleagues *Jubin Sebastian* and *Nidhal Mars* for their support and guidance which has always helped in local cultural integration and administrative steps. I would also like to express my gratitude for my elder brother *Numair Amjad*, who has always been very helpful in every aspect. These three years would not have been easy without his support.

An honorable mention goes to my beloved parents who have taken great care of me through thick and thin of my academic life and have been a great source of inspiration for me. I am

forever indebted to my mother and father for their unconditional love and endless prayers. No words can actually describe their everlasting love to me. I owe a lot to them; they encouraged and helped me at every walk of my life. Their unwavering faith and confidence in my abilities always motivated me.

Last but certainly not least, I wish to thank my beloved wife *Zainub* for her love, for sharing the ups and downs, and for her faith in me. My deepest love and respect to her for her tremendous patience, constant moral support, and endless prayers, which were invaluable in completing this journey. Special mention goes to my innocent, joyful and lovely son *Zayan* in whose company I forget all my worries.

Zubair Amjad
May 2020

Publications

Parts of this work have been published in the following journal and conference proceedings:

- **Z. Amjad**, A. Sikora, B. Hilt,& J.P. Lauffenburger, (2018, June). Low latency V2X applications and network requirements: Performance evaluation. In 2018 IEEE Intelligent Vehicles Symposium (IV) (pp. 220-225). IEEE.
- **Z. Amjad**, A. Sikora, J.P. Lauffenburger,& B. Hilt, (2018, August). Latency reduction in narrowband 4g lte networks. In 2018 15th International Symposium on Wireless Communication Systems (ISWCS) (pp. 1-5). IEEE.
- **Z. Amjad**, A. Sikora, B. Hilt,& J.P. Lauffenburger, (2018, September). Latency Reduction for Narrowband LTE with Semi-Persistent Scheduling. In 2018 IEEE 4th International Symposium on Wireless Systems within the International Conferences on Intelligent Data Acquisition and Advanced Computing Systems (IDAACS-SWS) (pp. 196-198). IEEE.
- K. A. Nsiah, **Z. Amjad**, A. Sikora, B. Hilt,& J.P. Lauffenburger, (2018). Performance Evaluation of Ultra-Low Latency Wireless Communication in Industrial Automation. In Embedded World 2018.
- K. A. Nsiah, **Z. Amjad**, A. Sikora, B. Hilt,& J.P. Lauffenburger, (2019, September). Latency Reduction Techniques for NB-IoT Networks. In 2019 10th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS) (Vol. 1, pp. 478-482). IEEE.
- K. A. Nsiah, **Z. Amjad**, A. Sikora, & B. Hilt, (2019, September). Performance Evaluation of Latency for NB-LTE Networks in Industrial Automation. In 2019 IEEE 30th Annual International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC) (pp. 1-7). IEEE.
- **Z. Amjad**, K. A. Nsiah, B. Hilt, J.P. Lauffenburger, & A. Sikora, "Latency Reduction for Narrowband URLLC Networks: A Performance Evaluation", Wireless Networks. (under second round of revision as of 6th April, 2020).

Table of Contents

List of Figures

List of Tables

Definitions

- **Machine Type Communication (MTC)**: Machine-type communications or machine-to-machine communications (M2M) refer to automated data communications among devices and the underlying data transport infrastructure. The data communications may occur between an MTC device and a server, or directly between two MTC devices.
- **End-to-End Latency**: Time it takes for a given piece of information to transfer from a source to a destination, measured at the communication interface (PHY OSI level), from the moment it is transmitted by the source to the moment it is successfully received at the destination. [1].
- **Control-Plane Latency**: Time it takes for a device to switch from idle to connected state [2]. It includes latency from synchronization signals and Random Access (RA) procedure.
- **User-Plane Latency**: Time between the data is available to protocol stack for transmission (APP OSI level) on the sender side and the data is available to be used after reception and processing on the receiver side (APP OSI level) [2].
- **Round Trip Time**: Time it takes for a signal to be sent plus the time it takes for an acknowledgment of that signal to be received [1]. This is usually measured at the PHY level.
- **Jitter**: It refers to the variation in end-to-end latency of packets carrying voice or video data over a communications channel.
- **Quality of Service**: Totality of characteristics of a telecommunications service that bears on its ability to satisfy stated and implied needs of the user of the service [3].
- **Reliability**: In the context of network layer packet transmissions, percentage value of the amount of sent network layer packets successfully delivered to a given system entity within the time constraint required by the targeted service, divided by the total number of sent network layer packets [1]. This is usually measured at the PHY layer.

Abbreviations

3GPP 3rd Generation Partnership Project

6LoWPAN IPv6 over Low-Power Wireless Personal Area Networks

AP Access Point

ARQ Automatic Repeat request

AS Access Stratum

BS Base Station

BSR Buffer Status Report

CCH Control Channel

CN Core Network

C-RNTI Cell Radio Network Temporary Identifier

CRS Common Reference Signals

CTTC Centre Tecnològic Telecomunicacions Catalunya

D2D Device-to-Device

DBPSK Differential Binary Phase-Shift Keying

DCI Data Control Indication

DL Downlink

DMRS Demodulation Reference Signals

DS Dynamic Scheduling

DSRC Dedicated Short Range Communication

eMBB enhanced Mobile BroadBand

eNB evolved Node Base station

EPC Evolved Packet Core

E-UTRAN Evolved UMTS Terrestrial Radio Access Network

FUA Fast Uplink Access

GFSK Gaussian Frequency Shift Keying

HARQ Hybrid ARQ

IIS Fraunhofer Institute for Integrated Circuits

IoE Internet of Everything

IoT Internet of Things

IP Internet Protocol

IPSec Internet Protocol Security

ISM Industrial Scientific and Medical

ITS Intelligent Transportation Systems

ITSA Intelligent Transportation Society of America

KPI Key Performance Indicators

LoRaWAN Long Range Wide Area Network

LPWAN Low Power Wide Area Network

LTE Long Term Evolution

M2M Machine-to-Machine

MAC Medium Access Control

MBB Mobile Broad-Band

MCS Modulation and Coding Scheme

MIB Master Information Block

MIOTY My IoT

MME Mobility Management Entity

mMTC massive Machine Type Communication

MTC Machine Type Communication

NAS Non-Access Stratum

NB-IoT Narrowband-IoT

NFC Near-Field communication

ns-3 Network Simulator

OBU On-Board Unit

PAN Personal Area Network

PDCP Packet Data Convergence Protocol

PDN Packet Data Network

PDU Packet Data Unit

P-GW Packet Data Network Gateway

PHY Physical

PRACH Physical Random Access Channel

PRB Physical Resource Block

PSS Primary Synchronization Signal

PUCCH Physical Uplink Control Channel

PUSCH Physical Uplink Shared Channel

QoS Quality of Service

RA Random Access

RAN Radio Access Network

RAR Random Access Response

RA-RNTI Random Access Radio Network Temporary Identifier

RBG Resource Block Groups

RLC Radio Link Control

ROHC Robust Header Compression

RRC Radio Resource Control

RSU Road Side Unit

SC-FDMA Single Carrier Frequency Division Multiple Access

SCHs Service Channels

SDU Session Data Unit

SG Scheduling Grant

S-GW Serving Gateway

SIB System Information Block

simuLTE simulator for LTE networks

SPS Semi-Persistent Scheduling

SR Scheduling Request

SRS Sounding Reference Signal

SSS Secondary Synchronization Signal

sTTI short Transmission Time Interval

TB Transport Block

TCP Transmission Control Protocol

TDMA Time Division Multiple Access

TTI Transmission Time Interval

UDP User Datagram Protocol

UE User Equipment

UNB Ultra Narrow Band

URLLC Ultra-Reliable Low-Latency Communication

USIM Universal Subscriber Identity Module

V2V Vehicle-to-Vehicle

V2X Vehicle-to-Anything

VoLTE Voice over LTE

WAVE Wireless Access for Vehicular Environment

1

Introduction

1.1 Résumé

La dernière décennie a vu l'émergence des technologies de communication sans fil par rapport aux communications filaires. L'un des moteurs de cette évolution technologique est l'internet des objets (IdO - Internet of Things/IoT). Cependant, les domaines d'application visés aujourd'hui par les communications sans fil vont largement au-delà des objets communicants de l'IdO et conduisent à l'Internet de tout (IdT - Internet of Everything/IoE). Ce nouveau paradigme de communication repose sur une grande variété de technologies sans fil, allant des technologies à très courte portée (quelques mètres) aux technologies à longue portée (plusieurs dizaines de kilomètres). Les réseaux sans fil offrent de nombreux avantages par rapport aux réseaux filaires, tels que l'évolutivité, la flexibilité, des coûts de déploiement réduits, etc. Toutefois, les réseaux sans fil ont à relever un certain nombre de défis pour devenir un choix privilégié pour les communications de machine à machine (Machine Type Communication/MTC). Ces défis comprennent, entre autres, la réduction des temps de latence, la gestion de l'extensibilité et de la densité des appareils.

Les réseaux cellulaires plébiscités en raison des nombreux avantages qu'ils offrent par rapport à d'autres types de réseaux à longue portée. Le premier chapitre de cette thèse présente un aperçu du fonctionnement des réseaux cellulaires ainsi que les défis à relever pour prendre en charge les applications de l'IdO. Une introduction au MTC et à ses exigences sont également présentées dans ce chapitre. Sur la base de leurs contraintes pour les réseaux cellulaires, un aperçu détaillé de cas d'utilisation nécessitant de très faibles latences est également présenté. Les cas d'utilisation sélectionnés concernent les Systèmes de Transport Intelligents (STI – Intelligent Transportation Systems/ITS) (STI) et l'automatisation industrielle. Ensuite, une

introduction à la simulation des réseau informatiques est présentée ainsi que les simulateurs open-source qui offrent la possibilité de simuler des réseaux cellulaires.

Les sections suivantes de ce chapitre mettent également en évidence les questions de recherche issues d'une étude bibliographique présentée dans le chapitre 4 de cette thèse. Ces questions de recherche portent non seulement sur les questions habituelles liées aux normes de réseau cellulaire, mais aussi certaines questions plus inhabituelles. Ainsi, les objectifs de ce travail de thèse issus de ces questionnements scientifiques de recherche sont également présentés. Ce chapitre se finit par un bref aperçu des contributions et de l'organisation de la thèse.

1.2 Introduction

The last decade has seen the emergence of wireless technologies over wired communication. One of the drivers of this technological change is the Internet of Things (IoT). However, the application areas targeted nowadays by wireless communications go largely beyond the IoT communicating objects and lead to the Internet of Everything (IoE). This new communication paradigm is based on a large variety of wireless technologies, ranging from very short to wide range technologies. Wireless networks offer many advantages over wired networks such as, scalability, flexibility, lower deployment costs etc. Numerous wireless communication technologies are available nowadays, ranging from very short range of a few meters to wide range of tens of kilometers. However, there is a number of challenges ahead of wireless networks to become a preferred choice for Machine Type Communication (MTC) [4]. These challenges include but are not limited to latency reduction, scalability, and device density.

Cellular networks have been in the highlight due to many benefits they offer over other wide area networks, such as already available infrastructure, wide range and the support of a large variety of device categories. In the following subsections of this chapter, an overview of cellular networks is given along with the use cases requirements that these networks should fulfill to support MTC. The subsequent sections of this chapter also highlight the research questions, objectives, and contributions of this thesis.

1.3 Cellular Networks

Cellular networks are radio networks distributed over land through large cells where each one includes a fixed location transceiver known as Base Station (BS) and User Equipment (UE) that are potentially mobile. These cells together provide radio coverage over larger geographical areas. A UE, such as a mobile phone, is therefore able to communicate even if the equipment is moving through cells during transmission [5]. The cellular networks have been evolving from first generation (1G) to current fourth generation (4G) and upcoming fifth generation (5G) in the last three decades mainly because of the user-centric Mobile Broad-Band (MBB) applications that demand higher data rate, as shown in Figure 1.1. Such applications normally focus on providing high data rates to the end users who need different services on mobile devices (e.g. cell phones). Due to the increasing demand around IoT applications, cellular networks are expected to support not only MBB applications but also MTC [6].

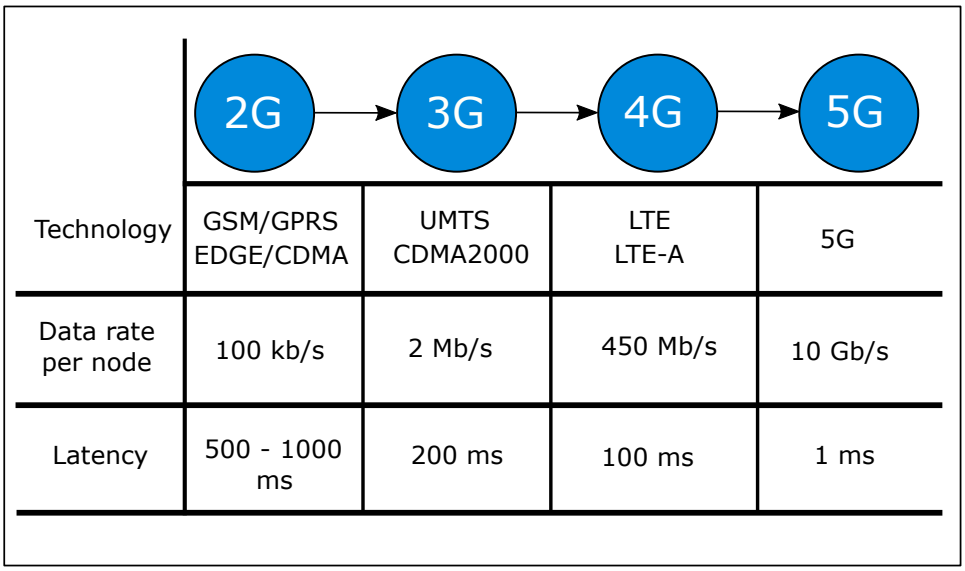

Figure 1.1: The evolution of latency and data rates in cellular networks form 2G to 5G. From 2G to 4G, data rate was the major factor in the development of the next generations; however, the current demand of low latency has also been driving the evolution from 4G to 5G, figure adapted from [7].

MTC is a type of communication that allows non-human handled devices to communicate with each other. This includes for instance IoT sensors, traffic lights, monitoring and control of industrial applications etc. There are two categories of MTC use cases named as Ultra-Reliable Low-Latency Communication (URLLC) and massive Machine Type Communication (mMTC). Both type of use cases are described in detail in section 1.4. Some examples of applications and devices from cellular networks for these cases are listed in the Figure 1.2. The major historical factor in the evolution of cellular networks was data rate increase. However, the current evolution of cellular networks is also being driven by the demands of low latency and massive device density.

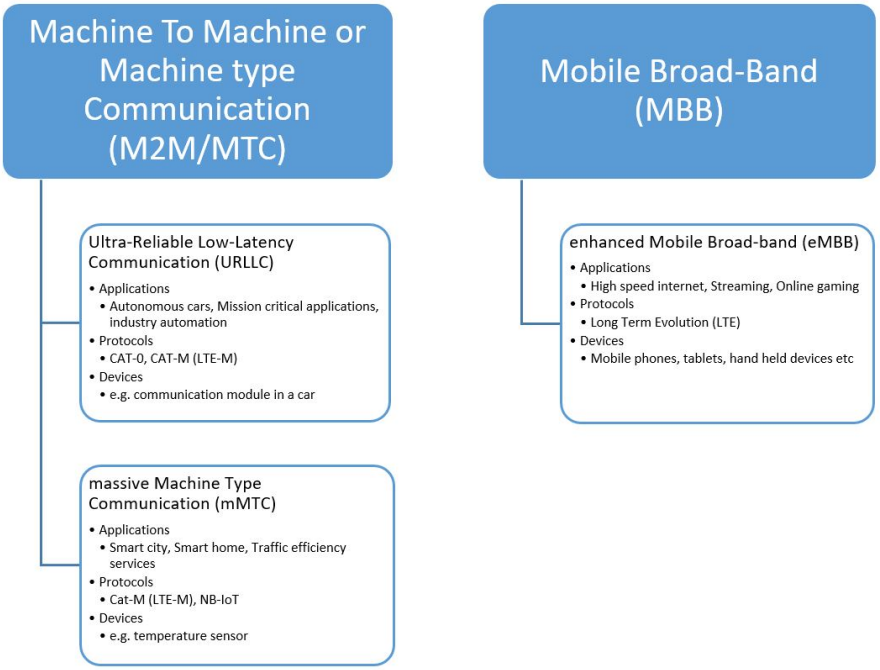

Figure 1.2: Mobile Broad-Band and Machine Type Communication use cases. The cellular networks are expected to support MTC use cases.

The cellular networks have been continuously improved and standardized by the 3rd Generation Partnership Project (3GPP) through different releases [8]. Figure 1.3 shows the generic LTE network architecture. There are two parts of the network called Radio Access Network (RAN) or Universal Terrestrial RAN (UTRAN) and Evolved Packet Core (EPC). The wireless communication takes place between UE and evolved Node Base station (eNB) in the RAN side. The EPC is responsible for network management. The major improvements until 3GPP Release 13 were focused on MBB applications. Therefore, the LTE latency remained unchanged from 3GPP Release 8 till Release 13 and does not meet the requirements of URLLC use cases. The latter require very low latency from the network with high reliability. For instance, an emergency control application requires the network latency to be less than 10 ms, which is not achievable by LTE RAN [9]. Additionally, 3GPP in Release 13 standardized two narrowband UE categories known as Cat-M1 (LTE-M) and Narrowband-IoT (NB-IoT) [10]. These UE categories were included in LTE specifications to support the narrowband cellular communication, which provides better coverage, longer battery life and lower manufacturing cost for devices. Due to these benefits, LTE-M and NB-IoT are considered good candidates for MTC.

One of the most important characteristics for MTC is the latency. There are multiple components that contribute to the latency in LTE networks. These components have a direct impact on the performance of the system. In Chapter 3 of this thesis, an in-depth latency

analysis for 4G LTE network latency is presented with a discussion towards the need for improvements for URLLC applications.

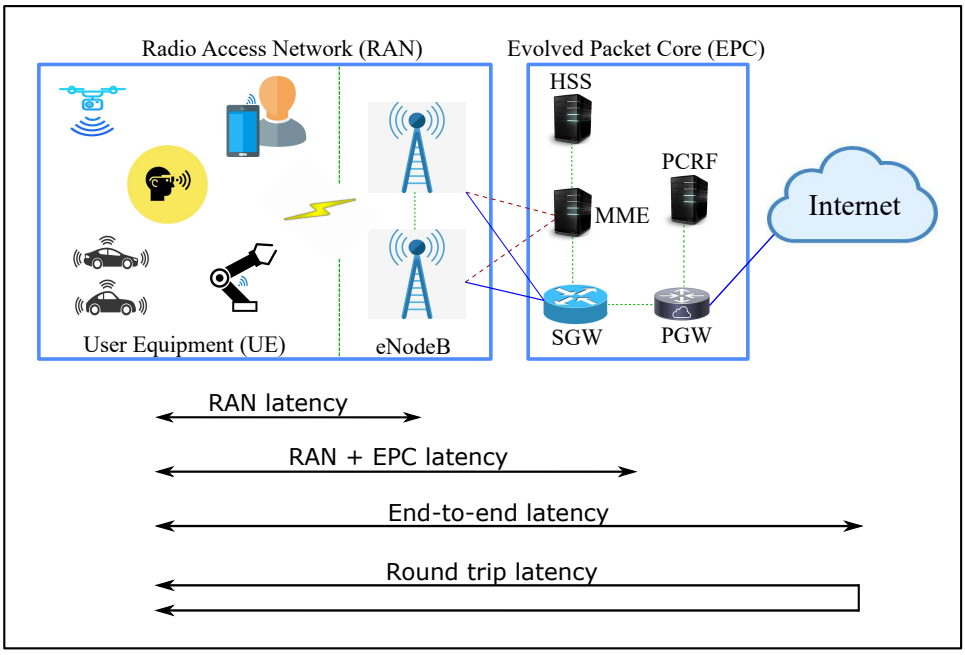

Figure 1.3: A generic architecture of the 4G LTE cellular network, adapted from [11]. The wireless communication takes place in the radio access network (RAN) while the network management is done by the evolved packet core. The description of the network entities is given in Chapter 2.

1.4 Machine Type Communication

MTC or Machine-to-Machine (M2M) communication enables the information exchange between electronic devices, including sensors, actuators, electronic appliances usually called connected Things, and/or MTC server. Communications can use both wireless and/or wired networks. MTC enables an endless number of applications in a wide plethora of domains, impacting different environments and markets [12]. It connects a number of devices to the Internet and other networks, forming the IoT. Several forecasts state a significant market growth over the next few years for both MTC devices and MTC connectivity segments [13, 14]. According to these forecasts, billions of machines or industrial devices will be potentially able to benefit from MTC, forming the massive MTC use case category for cellular networks (see Fig. 1.4).

One of the most important areas and challenges for the communication technologies is to provide support of wireless communication for MTC [15]. We see an accelerating use of connected devices around us and what we see now is still just the beginning. The vast majority of humans are already connected with smart phones. But we also use connected tablets, cameras that automatically upload pictures to the cloud, or use a remote control to supervise the summer house. We see more uses of simple wireless sensors that can keep track of temperature and rainfall, provide tracking of cargo containers, or measure water

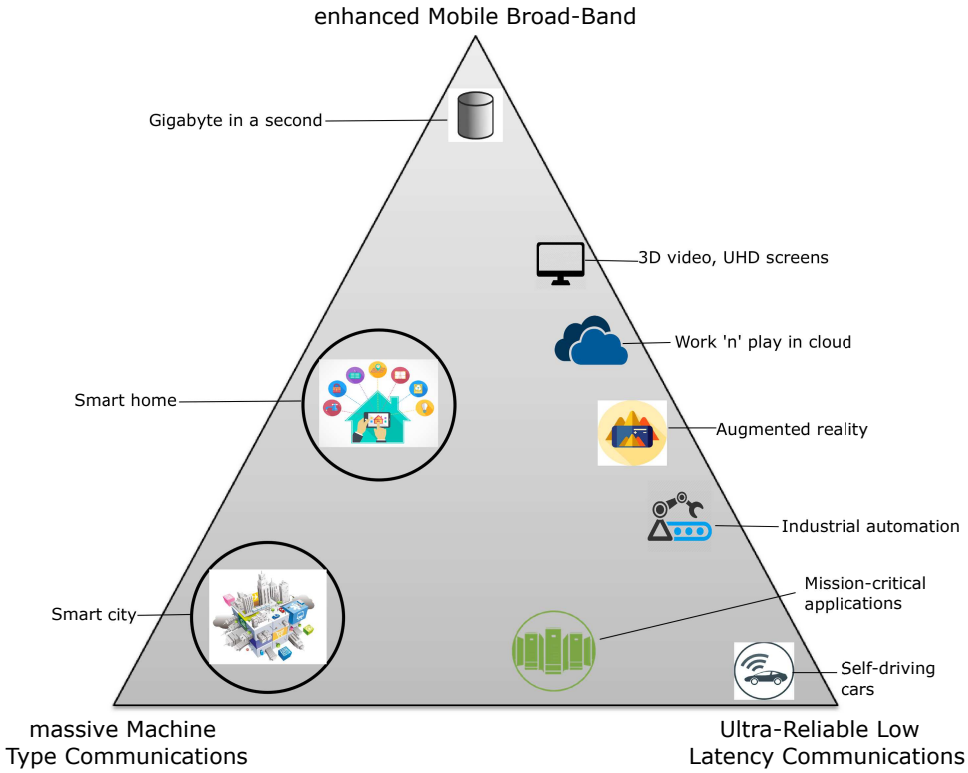

Figure 1.4: Three categories of cellular communication use cases as presented by the ITU in IMT 2020, figure adapted from [16].

or electricity consumption. Most of the time, transferring real-time information, down to a fraction of a second, is not required. Now, in the field of IoT MTC, many devices are being powered by batteries or solar cells that can seldom-–or never-–be recharged. Therefore battery consumption becomes as crucial. The major requirement from such applications is the coverage, which should be fulfilled by the wireless communication technologies.

Apart from above mentioned requirements such as energy consumption and data rate, another kinds of MTC have very different requirements. For instance, Intelligent Transportation Systems (ITS) will provide vehicles with information about road work or accidents ahead. The information may come from fixed base stations alongside each of the roads or from other, fellow vehicles, traveling in a near area. Vehicles normally come with lots of electrical power, so communication power efficiency is perhaps not the most important characteristic. The challenge is that the latency needs to be very short in order to be efficient for alerting the approaching hazards. Yet another use of communicating devices is industrial applications. In this particular field, communication tools can be used for controlling heavy machinery in remote or hazardous places; or used for monitoring and control of smart grids. Here providing extremely fast and reliable connectivity is the goal of wireless communication technologies.

It is obvious that the requirements for communication technology are very diverse, depending on the use case and type of device. Therefore, next generation cellular networks are being developed around the requirements of three use cases categories as presented in Figure 1.4 [16].

The top corner of the use case triangle represents the enhanced Mobile BroadBand (eMBB), which is targeted towards human users for providing higher data rates. The use cases in the lower left region is defined as the mMTC and require a very large number of devices supported by the network. The third type of use cases fall under URLLC in the right bottom that demand very low latency from the network.

These three use case categories are briefly explained in the following. A detailed discussion around the requirements and challenges of URLLC use cases is presented in Section 1.5.

- **enhanced Mobile Broad-Band (eMBB)**: High bandwidth Internet access suitable for web browsing, video streaming, and virtual reality. This is the Internet access service we use with smart-phones. Peak data rate of 10 to 20 Gbps, high mobility support of up to 500 kmph, macro and micro cell support, and reduced energy consumption are the highlighted features supported by eMBB.
- **massive Machine Type Communication (mMTC)**: Narrowband Internet access for sensing, metering, and monitoring devices. mMTC involves a large number of low-cost devices with high requirements on battery life and scalability. Possible scenarios for mMTCs include but are not limited to structure and environmental monitoring, asset tracking, process monitoring and optimization, maintenance of road signs and lightening, and smart waste disposal. High device density, long range, lower data rates, lower device cost, and longer battery life are the features supported by mMTC.
- **Ultra-Reliable Low-Latency Communication (URLLC)**: Services for latency-sensitive devices for applications like factory automation, cooperative driving, and remote surgery. URLLC targets mission critical applications with stringent requirements on latency and reliability [17–19]. Less than 1 ms air interface latency, low to medium data rates, high speed mobility, and ultra-reliability are the major features supported by URLLC.

Low latency allows a network to be optimized for processing data with minimum delay. The networks need to adapt to a broad amount of changing data in real time. URLLC is arguably the most promising addition to the cellular network capabilities. It requires a Quality of Service (QoS) totally different from MBB services. QoS is the ability to provide different priorities to different applications, users, or data flows, or to guarantee a certain level of performance to a data flow. URLLC provides networks with instantaneous and intelligent systems, though it requires transitioning out of the core network. On the other hand, mMTC require support for a very high device density. With the inclusion of mMTC, the number of connected devices in a network base station is expected to rise more than 100 times. Along with

this, a larger network range is also a challenge that the cellular networks have to fulfill. Figure 1.5 [20] presents a mapping of MTC and URLLC use cases on the latency and reliability canvas. There are use cases that do not require very low latency and can be supported by 4G cellular networks, such as process automation where different sensors and actuators installed in a factory hall are connected to the network and send the diagnostics and monitoring information to a central server. However, there are also applications, for instance factory automation where the control units are driven by the network or ITS that require the network to support low or very low latency. Therefore, to support such latency-sensitive use cases, improving the latency in those networks is of utmost importance. Furthermore, the device density required for factory automation is higher than that of ITS. The network is therefore required to provide scalability as well.

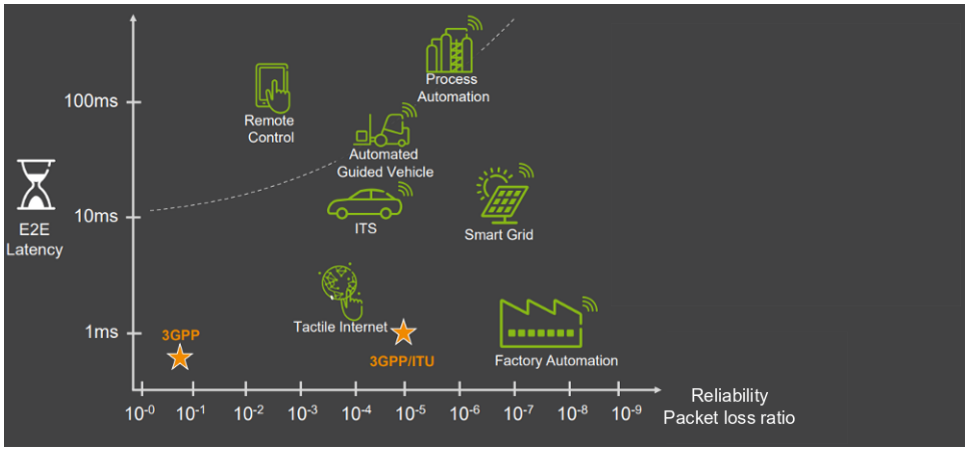

Figure 1.5: Reliability and latency requirements for different use cases that 5G networks should fulfill according to [20].

1.5 Selected Use Cases and Requirements: General Presentation

The evolution of cellular networks towards 5G has to take into account a variety of novel use cases that require more traffic [21] while maintaining high reliability and low latency as depicted above. The use cases of 5G communication focus on three major requirements as following.

- Higher data rate: this is especially the case when considering Mobile Broad-Band communication and cell phone communication including web surfing, streaming etc.,
- Higher device density: this is the case for massive Machine Type Communication (mMTC),

- Ultra-low latency: it is the most important requirement for Ultra-Reliable Low-Latency Communication (URLLC).

In the domains of control automation and automotive, for control processes, the very low uplink latency is one of the most critical requirements. Current 4G cellular networks have a nominal latency of about 50 ms to 100 ms including control- and user-plane latencies. However, this is currently unpredictable and can go up to several seconds [22] depending on many factors, that include but are not limited to the random access procedure, the device density, and the available system bandwidth. Moreover, 4G networks are mainly optimized for MBB traffic for very high data rates.

In order to set an applicative framework for our work and to be able to carry out realistic evaluations, in the following subsections we introduce two emerging mission-critical use cases, industry automation and ITS. Three applications of ITS/V2X and two for industry automation (see Table 1.1) are considered in this thesis and characterized on the basis of different requirements as summarized in Table 1.2. Both use cases and their respective applications are selected by considering their requirements for latency, energy consumption, device cost, device density, and communication range.

Table 1.1: Two use cases of URLLC and their applications considered in this thesis. Both use cases are selected based on the research areas of both partner institutes during this work. Each of the use case is further divided into applications based on mainly the requirements of latency, device density, and communication range.

Use case	**Application**	
ITS/V2X	Collision warning	
	Autonomous driving	
	Traffic efficiency services	
Industry automation	Factory automation	Monitoring
		Emergency control
		Closed loop control
	Process automation	

Even if in this thesis the scope is put on 5 different applications related to ITS and automation, it has been observed that most of the applications from these use cases are potential candidates for using narrowband cellular networks [23] meaning they do not require a very high bandwidth but rather involve thousands of devices and pertain to mMTC use cases or require low latency and pertain to URLLC use cases. That's why, in the upcoming

sections, that show the details of the ITS and automation use cases, we indicate for each one its belonging to one of these category of use cases.

1.5.1 Intelligent Transportation System / Vehicle to X Communication

ITS aims to provide improved movement, safety and journey experience by using information technology, to allow communication between vehicles (V2V) or between vehicles and infrastructure (V2I) [24]. The automotive market is evolving towards fully connected cars, which on one hand improves the user driving experience, but also is an essential requirement in cooperative driving, assisted overtaking, collision warning, and traffic efficiency services. Several use cases of ITS/V2X are under consideration in general and few of them fall under URLLC with stringent latency requirements. The latency-critical use cases of V2X based on different requirements are categorized and summarized in Table 1.2. For each use case, the requirements of different network characteristics are also mentioned. An illustration of different ITS/V2X applications is given in Figure 1.6. Three main categories of latency critical V2X applications are as follows.

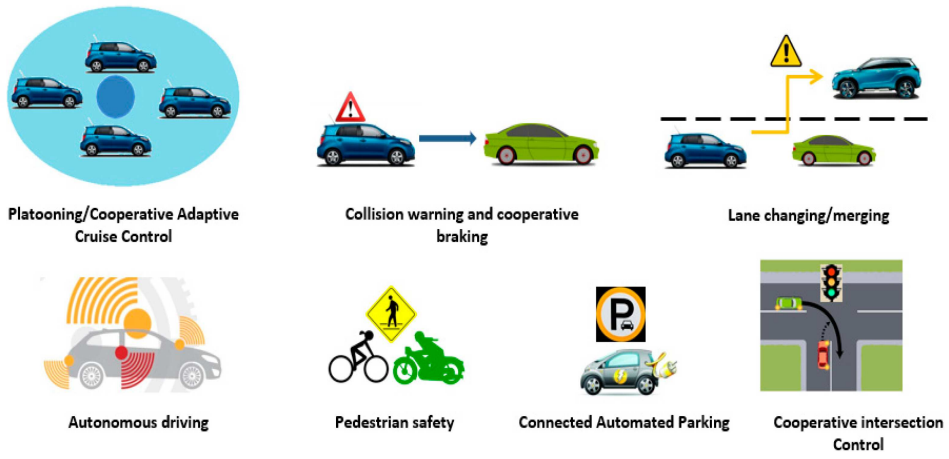

Figure 1.6: Different applications of ITS/V2X, according to [25].

1.5.1.1 Collision Warning - URLLC

Collision warning systems provide warning notifications to vehicles that are in hazardous conditions. Collision notification becomes most important when all other means of avoiding an accident fail. Vehicles can control their speed, direction, and acceleration to avoid collision on the basis of these messages. The latency requirements of collision warning vary, depending on the driving scenarios such as highway or urban traffic and are comprised between 10 and 100 ms [26]. A collision warning message can be transmitted from the base station to a vehicle or between vehicles. In both cases, the latency requirement needs to be fulfilled by the network.

The maximum number of devices per network cell could reach up to 300 in an urban area and 50 on a highway segment. The communication range of more than 2000 m is also required from the network to support collision warning.

1.5.1.2 Traffic Efficiency - mMTC

In the majority of traffic efficiency-oriented use cases, vehicles and other elements on the road either upload their information (e.g. position, speed, acceleration) or event information (e.g. road condition and traffic situation). These data transmissions are usually periodic and small sized. The data from these transmissions could be used in an extended way to support more complex and novel services and applications [26], such as see-through, bird's eye view for intersections, or vulnerable road user discovery. These services also require latencies less than 100 ms, as mentioned in Table 1.2. Similar to collision warning, these services also requires a range of more than 2000 m and a device density of 300 per network cell.

1.5.1.3 Autonomous Driving - URLLC

[26] defines six levels of driving automation depending on the amount of assistance an automated device provides to the vehicle. Level-0 only assists the human driver who keeps total control of the car while the level-5 allows driver-less fully autonomous vehicles. Figure 1.7 shows an illustration of these six levels (level-0 to level-5). The increase in levels is determined by the features offered by the vehicle toward autonomous driving. Level-0 offers no active assistance to the driver who is in charge of the total car control. In level-1 and level-2, the driver is required to steer the vehicle with complete attention, however, many different drive assist systems offer services to the driver, for instance active traffic information. The level-3 of automation requires the driver to take-over the control of the vehicle in order to perform an overtake. In level-4, the driver is not required to focus on the road or to control the car. The vehicles with level-5 automation can drive without any driver and therefore do not need any human intervention. The wireless communication can be part of an automated vehicle starting from level-1 in order to enable different applications and services. The low latency communication becomes important starting from level-4 onwards where the vehicle mostly drives itself without any input from the driver. Vehicles can benefit from the information coming both from other road users or from the network. The received information then helps vehicles to adapt to the traffic and road conditions. It has been shown in the literature that the cellular communication system required for such purposes needs to support a 10 ms higher bounded latency [26] while ensuring high reliability (i.e. 10^{-5} packet loss ratio) requirements.

The device density and range requirements for autonomous driving is similar to collision warning and traffic efficiency services.

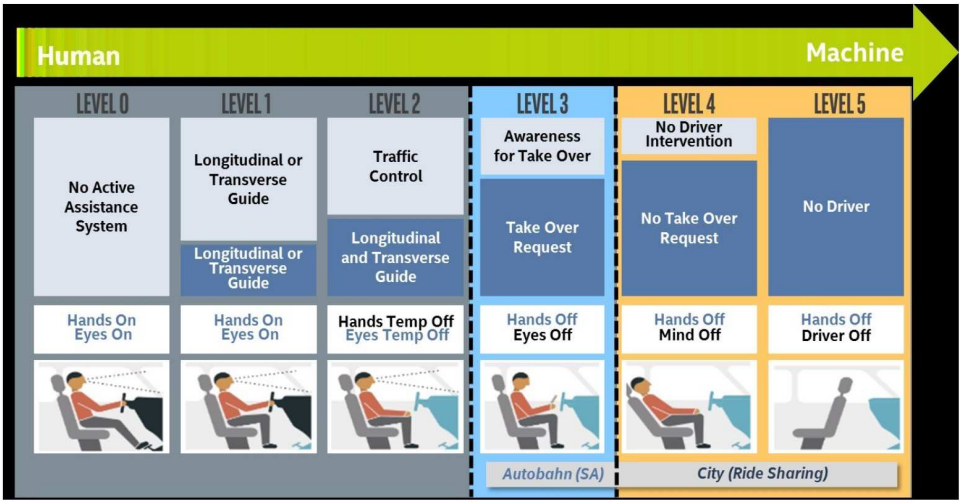

Figure 1.7: Six levels of automation in ITS. Each level defines its requirements, according to [27].

In this section we observed that the most stringent constraints for ITS communications are in the areas of autonomous driving and collision warning, where the maximum latency expected is 10 ms for the former and 100 ms for the latter.

1.5.2 Industry Automation

As shown in Figure 1.5 with the location of factory automation use case on the reliability and latency canvas, URLLC is one of the enabling technologies in Industry 4.0 [28]. In this vision, industrial control applications are automated by wireless network deployment in factories. Typical industry automation applications include factory automation and process automation. While factory automation includes applications that control the operation of production machines, process automation provides monitoring and diagnostics of machines. Traditionally, industrial control applications are based on wired networks as the current wireless networks cannot meet the latency requirements. Nevertheless, wireless technologies can bring substantial benefits including reduced deployment cost and maintenance, long-term reliability, and deployment flexibility. Both applications are described in detail in the following subsections.

1.5.2.1 Factory Automation - URLLC

Factory automation is defined as the use of control systems for operation of machinery and processes. It includes real-time traffic from sensors installed in machines (e.g. a high speed assembly for the control of machines) and is further categorized in monitoring and emergency control of machines and close loop control applications. Factory automation applications are

generally considered to be highly sensitive in terms of reliability and latency as presented in Table 1.2. These applications require latency from 1 ms to 50 ms with very high reliability (i.e. from 10^{-5} to 10^{-9} packet loss ratio) [28].

1.5.2.2 Process Automation - mMTC

Process automation includes the periodic traffic from sensors for monitoring and diagnostics of industrial elements, such as cooling, heating, mixing and so on. The latency requirements for process automation are relaxed as the change in measured values is relatively slower. Therefore, the latency requirements for such services range from 50 ms to 100 ms with a packet loss ratio of 10^{-3}. The coverage area for process automation is relatively large and usually includes multiple buildings and outdoor sites as well. Moreover, process automation also requires the communication network to support up to 300 devices in a cell. With such higher device density, it becomes a challenge for the network to provide guaranteed latency.

1.5.3 Selected Use Cases: Summary

To enable URLLC, next generation cellular networks have to provide very low latency, while supporting simultaneously a high device density and a very high reliability. Among other use case categories, URLLC requires guaranteed low latency from the network while keeping other requirements such as, reliability, range, device density, and mobility fulfilled. In this sub-section, a detailed introduction into two of URLLC use cases (Table 1.1) is provided. The latency requirements of use cases are also presented in Table 1.2 and discussed in detail. It is essential for the cellular networks to support these requirements in order to enable these use cases.

The work in this thesis is part of a collaboration project between IRIMAS lab [29] in Université de Haute-Alsace, France and Institute of Reliable Embedded Systems and Communication Electronics (ivESK) [30] in Offenburg University of Applied Sciences, Germany. The IRIMAS lab in France has active research in the field of control and signal processing for autonomous vehicles in the sub-group Modeling and Identification in Automatic and Mechanical Engineering (MIAM). Another group Network and Telecommunications (RT) inside IRIMAS has an active research in wired and wireless communication networks. The selection of V2X use case in this thesis is part of common research interests from these two groups in IRIMAS. From the German side, the partner institute ivESK at Offenburg University is active in research around embedded systems architecture, protocols and security, low power wide area networks, and narrowband cellular communication in the context of Industry 4.0. To align the common research interests, the industry automation use case is listed as one

of the selected URLLC use case. Both the selected use cases present mMTC and URLLC applications. The improvements made for these use cases also hold true for other use cases in the same category (i.e. mMTC and URLLC).

Table 1.2: Requirements for latency-critical use cases of MTC. The communication traffic type, depending on the application, can be either periodic or event triggered. [26, 28, 31, 32]

Use Case	**Application**		**Latency (ms)**	**Reliability (PLR)**	**Max. number of devices per cell**	**Comm. range (m)**	**% of mobile devices**	**Mobility speed (km/h)**	**Traffic type**
ITS/V2X	Collision warning (URLLC)		10 to 100	10^{-3} to 10^{-5}	Urban-300, Highway-50	Urban-500, Highway-2000	>90	Urban<100, Highway<500	ET
	Autonomous driving (URLLC)		10	10^{-5}	Urban-300, Highway-50	Urban-500, Highway-2000	>95	Urban<100, Highway<500	ET
	Traffic efficiency (mMTC)		<100	10^{-3}	300	2000	>80	<500	Periodic
Industry Automation	Factory automation (URLLC)	Monitoring	<50	10^{-5}	100	100 to 500	0	0	Periodic
		Emergency control	<10	10^{-9}	50	50 to 200	<5	<20	ET
		Closed loop control	<10	10^{-9}	100	50 to 200	0	0	Periodic
	Process automation (mMTC)		50 to 100	10^{-3}	300	100 to 2000	<5	<5	Periodic

*ET: Event Triggered, Latency: The delay between UE and eNB, PLR: Packet Loss Ratio, URLLC: Ultra-Reliable Low-Latency Communication, mMTC: massive Machine Type Communication

1.6 Network Simulators

In order to investigate and develop different techniques for improving wireless networks, especially in the research field, use of network simulators is very essential. Network simulators are used to investigate and evaluate different protocols, which are difficult in terms of resources to be performed in real life. For instance, to evaluate the behavior of a large number of devices in a cellular network without a simulator, one must have access to the infrastructure of a network operator, which is usually not the case for academic researchers. In such situations, a simulator provides a good blend of resources and abstraction of real life scenarios. Network simulation is performed by a software program simulating the interaction of devices and equipment. The aim of a simulator is to test the end-to-end application behavior on different network designs. The network simulation is completely performed by software.

There is a number of open-source simulators available for the research community to not only simulate the networks of their choice but also to contribute in the developments of different features offered by the simulators. This section, without having the objective of being exhaustive, provides a glimpse of view of the currently available solutions. Some examples of simulators available for cellular networks are Network Simulator (ns-3) [33], simulator for LTE networks (simuLTE) [34], which is based on OMNet++ [35], and Matlab [36]. The Matlab based simulator supports simulation of mainly the Physical (PHY) layer and channel characteristics. It does not support the core network simulation. simuLTE is based on the OMNet++ simulator supports core and radio access network simulation. The LTE protocol stack for devices and base stations is implemented only for the user-plane messages and simulation of control-plane messages is not supported. Among these simulators, ns-3 offers the most comprehensive implementation of LTE protocol stack for RAN and Core Network (CN) parts of cellular networks. It can also be noticed that for some simulators, of which ns-3 is one, the term simulator may be outdated and the term emulator would apply even better to them. There are two main components of the LTE module in ns-3.

- **LTE model**: This model includes the LTE Radio Protocol stack. These entities reside entirely within the user devices and the base station nodes. This model stands for the UTRAN or RAN and will be mainly used for our work.
- **EPC model**: This model includes core network interfaces, protocols and entities. These entities and protocols reside within different components of EPC, and partially within the base station nodes.

ns-3 is a discrete event-driven simulator that allows the system model to evolve as a sequence of events, where an event represents a change in the model state. The ns-3 simulation core supports research on both IP and non-IP based networks. The LTE module of ns-3 offers a complete implementation of the radio protocol stack, core network entities and application support. Therefore, ns-3 was selected for the implementation and evaluation of proposals in this thesis. However, LTE module of ns-3 lacks the implementation of NB-IoT. Additional non technical reason for choosing ns-3 was the fact that the two partner labs of this PhD use currently this simulation tool in their works.

1.7 Research Questions and Contributions

This thesis tackles the research challenges around the latency and scalability of LTE networks and especially 4G networks for URLLC use cases. The approaches presented in this thesis towards the solution of 4G LTE latency problems are also relevant for 5G cellular networks as the use cases considered in this thesis are defined for 5G networks and as improved 4G networks would also be part of the 5G networks. More precisely, the following research questions are investigated:

- **Q1. What are the inherent limitations of LTE with respect to achievable latency?** It is essential to first understand the limitations of 4G LTE in terms of user-plane and control-plane latency to point out the potential problems that could be improved. This analysis will help in defining technical solutions helping to achieve the latency required by the use cases described in Section 1.5.
- **Q2. Which approaches are best suited to optimize latency of cellular networks in practice?** To optimize the latency for different LTE standards of MTC such as LTE-M, and NB-IoT, it is required to investigate potential techniques to meet the URLLC use case requirements and to include improved 4G standards in 5G.
- **Q3. How can the minimum latency be achieved as functions of system characteristics?** The requirements of communication differ for different use cases of URLLC as already shown earlier. It is essential to point out the limitations and conditions, under which the latency requirements shall be fulfilled.
- **Q4. Which beyond-the-standards approaches are potential candidates for improving network performance, and in particular latency?** To optimize the latency of cellular networks, different approaches shall be investigated even outside the 3GPP standards.

To overcome the challenges associated with the above research questions, the following objectives have been outlined for the contributions of this thesis:

- **Contribution 1.** Investigate the LTE control- and user-plane latencies with realistic simulations in the open source ns-3 simulator. (Q1, Ch. 3)
- **Contribution 2.** Investigate the potential latency reduction techniques by developing the concepts in the simulator and evaluating through realistic simulations. (Q2, Ch. 4)
- **Contribution 3.** Implement and evaluate latency reduction techniques for LTE-M in ns-3 for URLLC use cases. (Q3, Ch. 4)
- **Contribution 4.** Implement and evaluate the NB-IoT module in ns-3 to evaluate the proposed techniques for narrowband device categories of MTC. (Q2, Ch. 4)
- **Contribution 5.** Implement and evaluate latency reduction techniques for NB-IoT in ns-3. (Q3, Ch. 4)
- **Contribution 6.** Conceptualize, implement, and evaluate the Hybrid Scheduling technique for LTE-M. (Q4, Ch. 5)
- **Contribution 7.** Conceptualize, implement, and evaluate a Flexible TTI structure for LTE-M. (Q4, Ch. 5)

The defined objectives are listed in Figure 1.8 for each of the research question mentioned above.

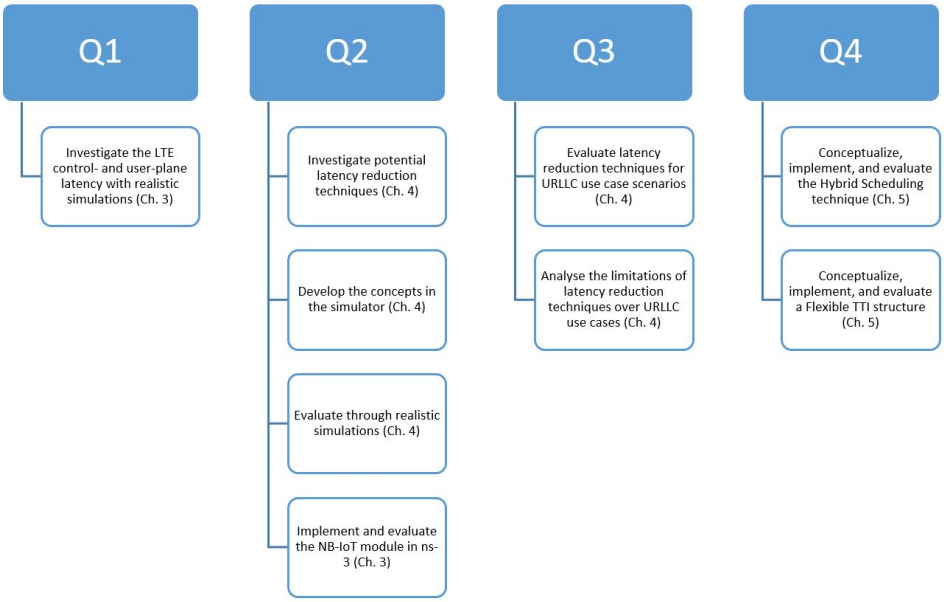

Figure 1.8: There are four research questions in this thesis that are raised from the literature survey. Each of the questions is approached with the defined objectives.

Figure 1.9 lists the ns-3 versions and respective contributions made by the author in this thesis. The respective chapters of the thesis are also indicated where each of the implementation was used for the simulations.

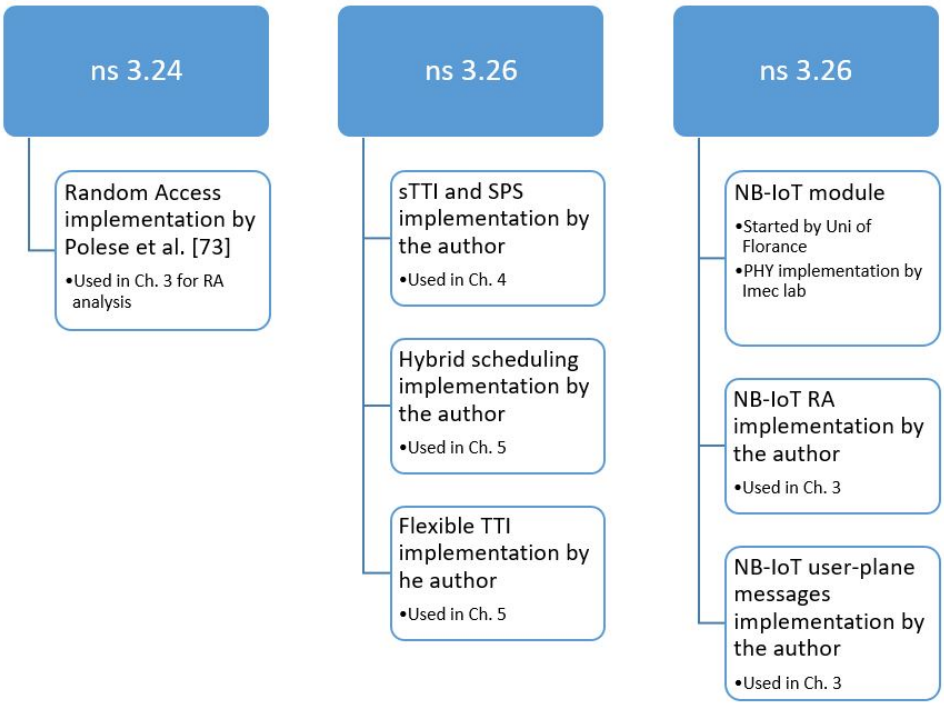

Figure 1.9: Different contributions used in this thesis. An already available model was used for simulation of Random Access. All other features and modules used in this thesis are the contributions of the author.

1.8 Thesis Organization

The core chapters of this thesis are structured as follows:

Chapter 2 provides an overview of the available wireless technologies for MTC. A detailed overview of 4G networks and their evolution towards 5G with a focus on MTC is also described. The narrowband devices standardized for MTC are also discussed. The architecture and protocol stack of LTE networks is also presented.

Chapter 3 presents an in-depth theoretical analysis of control- and user-plane latencies that are also evaluated with realistic simulations performed in an open-source simulator. The open-source ns-3 simulator design and architecture is also described.

Chapter 4 describes the design, literature survey, development and evaluation of latency reduction techniques. The simulations are performed with realistic URLLC use case scenarios. Moreover, a broad range of parameters are explored within the simulations.

Chapter 5 conceptualizes a hybrid scheduling scheme for efficient resource management and a flexible transmission time interval structure for diverse set of applications in cellular networks. The proposed schemes are developed and evaluated in ns-3.

Chapter 6 concludes the thesis with a summary of the work in this research, the discussion of possible future research directions, and final remarks.

2

Wireless Technologies for Low Latency Communication: State of the Art

2.1 Résumé

Les MTC couvrent un large éventail de cas d'utilisation qui vont d'un seul appareil avec des contraintes spécifiques à des déploiements massifs multiplateformes de technologies embarquées et de systèmes en cloud nécessitant des connexions en temps réel. E support de tout déploiement de l'IdO/TMC est le réseau lui-même. Dans les premiers déploiements MTC/M2M (Machine to Machine), les réseaux étaient câblés. Mais aujourd'hui ils sont de plus en plus souvent remplacés par des réseaux sans fil. Pour assurer une connectivité efficace aux dispositifs M2M, de nombreux protocoles sont disponibles, notamment ceux des réseaux personnels (Personal Area Networks/PAN) et les réseaux étendus de faible puissance (Low Power Wide Area Network/LPWAN). Les communications cellulaires sont normalement considérées comme un bon compromis entre portée et débit de données. Les réseaux cellulaires modernes offrent une grande variété de choix d'appareils, avec différents débits de données, portées et consommations d'énergie. Comme les travaux de cette thèse portent sur des cas d'utilisation de communications ultra-fiables à faible latence (Ultra-Reliable Low-Latency Communication/URLLC) qui sont sélectionnés sur la base d'une communication à faible puissance, à faible coût et à large portée, ce chapitre présente un aperçu détaillé de l'état de l'art des réseaux cellulaires, Wi-Fi et des LPWAN. Les réseaux cellulaires mobiles ont évolué au cours des trois dernières décennies en fonction des différentes exigences des applications dont ils véhiculent les données. Les réseaux LTE (Long Term Evolution) 4G ont été déployés et utilisés depuis un certain temps et sont

en train d'évoluer vers la cinquième génération de réseaux cellulaires. Il est prévu que les améliorations du LTE fassent également partie des réseaux 5G afin de réduire les coûts de déploiement. L'évolution vers les 5G et 6G est également abordée vers la fin de ce chapitre.

2.2 Introduction

Since the work in this thesis mainly revolves around low-latency communication use cases targeting factory automation and intelligent transportation systems (presented in section 1.5), it is important to first discuss the available wireless communication technologies. In doing so, this chapter first brings a discussion around recent developments in machine type communication. Afterwards, the available wireless technologies including Low Power Wide Area Networks, Wi-Fi, and cellular networks are discussed. Among these wireless technologies, based on the use case requirements, cellular networks have been selected as a potential technology to support the low latency use cases that potentially require long range and are also described in detail. Towards the end of the chapter, the recent developments of next generation cellular networks are presented.

2.3 Recent Developments in Machine Type Communication

The communication between machines covers a huge range of use cases that scale from a single constrained device such as a temperature sensor, up to massive cross-platform deployments of embedded technologies and cloud systems connecting in real-time. Figure 2.1 that stands for cellular MTC is used here in order to illustrate a general purpose case. It shows multiple domains of communication between machines architecture. The backbone of any machine communication deployment is the network itself. In earlier deployments, networks used to be wired, however nowadays it is shifting towards wireless networks.

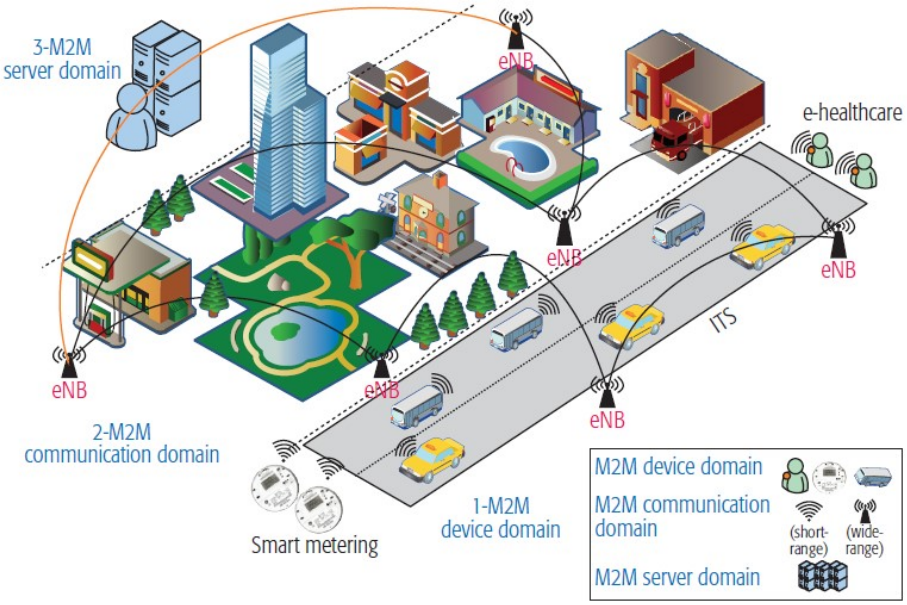

Figure 2.1: MTC communication architecture according to [37].

To provide efficient connectivity for machine type devices, many well defined protocols are available going from Personal Area Network (PAN) to Low Power Wide Area Network (LP-WAN). Examples of short range wireless PANs include Bluetooth, Wifi, Zigbee, WirelessHART and IPv6 over Low-Power Wireless Personal Area Networks (6LoWPAN). Figure 2.2 illustrates the mapping of different wireless communication technologies with respect to data rate and communication range. LPWANs provide long range while keeping the data rate as very low (a couple of bytes/s). Short range communication technologies like Bluetooth, Near-Field communication (NFC), etc. can provide higher data rates with a shorter communication range. However, cellular communication is normally considered to be a good trade-off between range and data rate. There is a wide variety of choices available for devices in modern day cellular networks, with different data rates, ranges, and energy consumption. Since the work in this thesis revolves around the URLLC use cases (described in Chapter 1) that are selected based

on low-power, low-cost, and wide-range communication, the following subsection presents LPWANs followed by a detailed introduction into cellular networks.

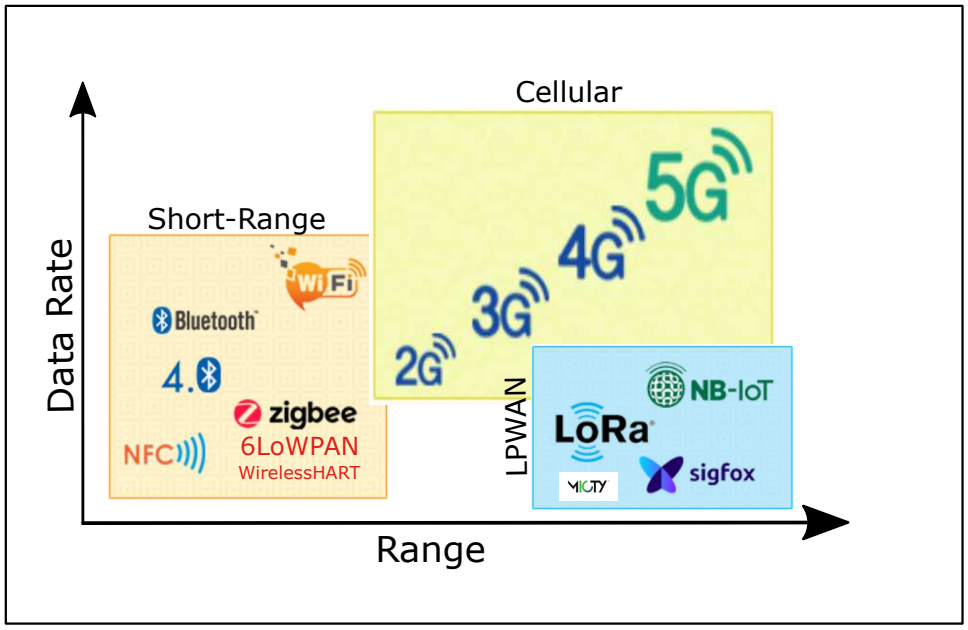

Figure 2.2: Heterogeneous wireless communication technologies, figure adapted from [38].

2.4 Low Power Wide Area Network Technologies (LPWAN)

One major challenge that IoT networks face is the fact that, as long as they transmit data, UEs such as sensors operated on battery have to stay charged. LPWAN, as the name already suggests, is a wireless technology with as major objective, among others, the maximization of the end device battery life while extending the coverage area. The power consumption of UEs in LPWAN is being minimized by:

- Allowing sleep times: Here, the main energy consumers, the radio transceiver and microcontroller, alternate between idle and active [39]. They use known power consumption patterns and timing constraints predefined by the Medium Access Control (MAC) layer.
- Bandwidth/data rate: Bandwidth is used to determine the amount of data being transferred (i.e., bit rate) usually per second. It is the spectrum range in Hertz that can be used by a system to transfer digital information (in digital communication) [40].

In a typical M2M network, the uplink (UL) traffic originates from the wireless end devices, destined to the Access Point (AP), the sink node, or the base station. This results from the fact that current M2M communication applications are mostly used for data gathering (going from the device to the network/cloud). The devices are generally designed for ultra-low-power operation as it is a key requirement to provide a battery lifetime of about 10 years or more [40].

There is a wide range of LPWAN technologies available. Among those, the most widely used and discussed in the following are Long Range Wide Area Network (LoRaWAN), Sigfox, and MIOTY [41]. These LPWAN technologies have very low data rates, lower power consumption and very high link budget. A link budget is an accounting of all the power gains and losses that a communication signal experiences.

2.4.1 Long Range Wide Area Network (LoRaWAN)

LoRa is a Low Power Wide Area (LPWA) networking standard between distributed sensor devices and distributed gateways [42]. LoRaWAN is a media access control layer protocol for managing communication between LPWA gateways and end-node devices, maintained by the LoRa Alliance. LoRaWAN defines the communication protocol and system architecture for the network while the LoRa physical layer enables the long-range communication link. LoRaWAN is also responsible for managing the communication frequencies, data rate, and power for all devices. LoRaWAN has an extremely low channel capacity with a low maximum data rate (<50 kbps), very low power consumption, and therefore a very high link budget. LoRaWAN uses the Industrial Scientific and Medical (ISM) radio bands on a license-free basis. Of the available ISM bands, LoRaWAN uses 683-870 MHz and 433 MHz in Europe and different bands in other parts of the world. A typical LoRaWAN network architecture is shown in Figure 2.3.

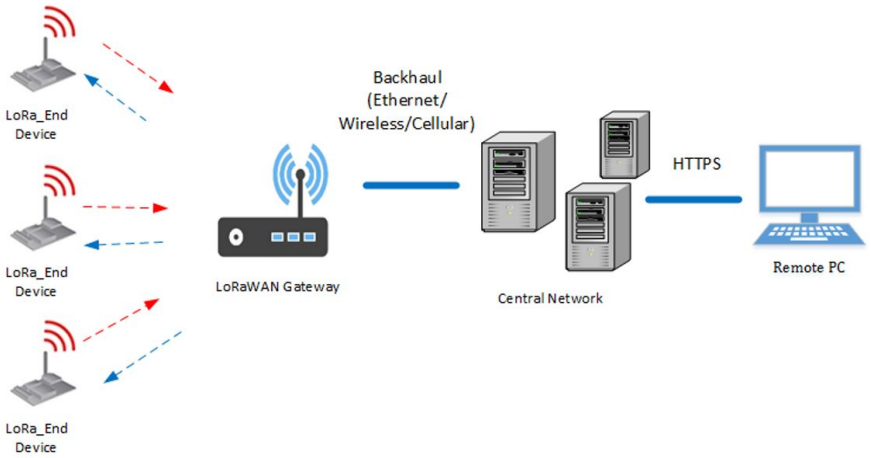

Figure 2.3: LoRaWAN network architecture.

It is laid out in a star topology, in which a device (an end device in the LoRaWAN terminology) is connected to a central server via gateways (or base stations) [43]. In this architecture, the LoRaWAN gateway acts as the transparent bridge relaying messages between the end devices and the central network in the back end. It also supports, for maximum battery saving, end devices with device-originated calls only. In addition, LoRaWAN supports a network-oriented transmission mode, in which end devices periodically wake up to receive

paging like LTE UE in a traditional cellular network. For paging to be facilitated, a beacon signal is transmitted by a gateway that allows an end device to synchronize to the network and look for downlink transmission in the predetermined windows [44].

2.4.2 Sigfox

Sigfox is an operator of a network fully dedicated to low-throughput communication for connected objects. With an extremely cost-effective and very low-energy consuming out-of-the-box connectivity offer, Sigfox brings a revolution to the world of IoT and M2M [45]. The network, which already connects tens of thousands of objects, is being rolled out worldwide. A typical Sigfox network architecture is shown in Figure 2.4. The Sigfox network has a star topology where end devices send messages when needed with bidirectional request to the Sigfox base stations/gateways within the range. All the Sigfox base stations deployed are directly connected to the Sigfox Cloud via a secure Internet Protocol (IP) connection. The base stations detect, demodulate, and report those messages to the Sigfox cloud. The Sigfox cloud then pushes the messages to the remote end user. The end user can also send a reply to the end devices over Sigfox cloud through the base stations.

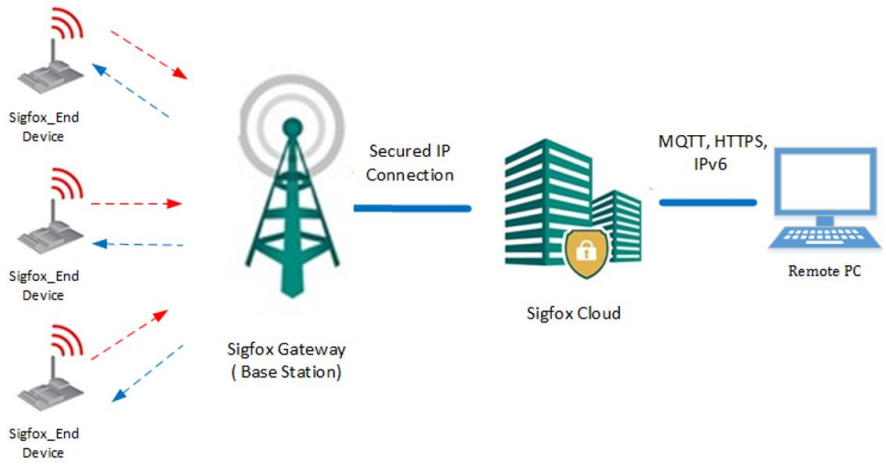

Figure 2.4: Sigfox network architecture [46].

Sigfox is a proprietary technology that employs the Differential Binary Phase-Shift Keying (DBPSK) and the Gaussian Frequency Shift Keying (GFSK) that operates in the 868 MHz ISM band. It transmits short messages (up to 12 bytes long) occasionally at low data rates (a node can send up to 140 messages per day). It provides a 162-dB link budget and long range of up to several kilometers. The end devices send messages when needed with bidirectional request to the Sigfox base stations/gateways within the range.

2.4.3 MIOTY

My IoT (MIOTY) is a LPWAN solution dedicated to private IoT Network developed by Fraunhofer Institute for Integrated Circuits (IIS) [47]. MIOTY is based on a Ultra Narrow Band (UNB) technology with very narrow signal bandwidth (2 kHz) to achieve long distance data communication (5 km in urban and up to 15 km in rural area) between thousands of IoT devices and a base station. A basic illustration of a MIOTY network is shown in Figure 2.5. The UNB technology provides maximum spectrum efficiency for best in class spectrum utilization. This solution relies on an asymmetrical transmission method that uses scores of simple sensor nodes and a complex receiver. MIOTY uses an efficient channel encoding scheme called telegram splitting that increases its range by a factor of 10 over standard 868 MHz wireless systems. In telegram splitting, a UNB signal is split into numerous smaller sub-packets and transmitted at different time and frequencies with transmission-free periods in between. Because it generates little self-interference, the system can support up to one million simultaneous transmitters. MIOTY is furthermore resistant to interference within its own frequency band. MIOTY can be deployed for a wide range of applications. The technology is suitable for monitoring large technical systems or difficult-to-access areas and can serve as the underlying transmission infrastructure for applications such as switching statuses, machine-to-machine communication, smart meters and other systems.

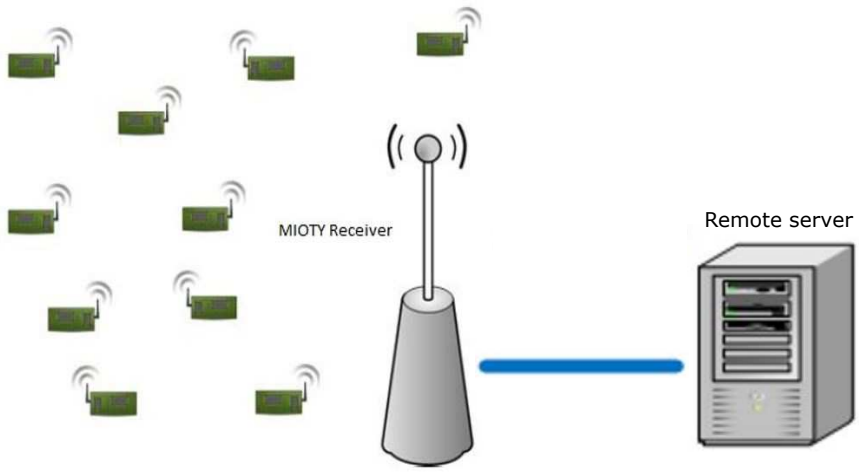

Figure 2.5: MIOTY network architecture.

The above mentioned LPWAN technologies provide very good long range communication with lower energy consumption. They are useful mainly in the mMTC use cases where the devices do not need very low latency and high data rate. The use cases considered in this thesis (presented in section 1.5) require low latency and in some cases high data rates such as ITS services. With these limitations of LPWANs, it is necessary to look for other alternatives

that provide a good blend of range, data rate, and latency. Therefore, in the following, some more technologies are discussed.

2.5 Wi-Fi Technologies

Wi-Fi allows networking of digital devices (e.g. computers, cell phones, tablets) without the need of wires. Wi-Fi uses a radio technology known as IEEE802.11 [48], which can transmit data over short distances using high frequencies. Based on IEEE802.11 standards, Wi-Fi operates on either 2.4 GHz or 5 GHz depending on its standard. The network's central point is the access point. Typically, the range of this Wi-Fi access point to any Wi-Fi capable devices is about 160 m outdoors and 70 m indoors. This estimated range does not take into account any obstructions, which may block the signal, including walls, solid objects or trees. The more obstructions in the signal's path from the base station, the shorter the range. The major advantages of Wi-Fi standards family are that they forms the staple of home, business and office networking and is widely used for its high data transfer rate abilities (max throughput of up to 54 Mbit/s with 12 Mbit/s being typical). However, complying with the standard, requires excessive overhead in terms of power consumption, software, processor resources, short ranges (160m max) and size of physical components, making it less effective in most industrial situations.

2.5.1 Wi-Fi for Industrial Applications

Industrial environments are uniquely different from office and home environments. High temperatures, excessive airborne particulates, multiple obstacles and long distances separating equipments, are special challenges that make it difficult to place and reach sensors, transmitters, and other data communication devices. These factors create a very unique, complex, and costly challenge for establishing data communication channels that are reliable, long lasting, and cost effective.

In an effort to integrate industrial systems with the other office devices, PROFINET has offered a protocol for an industrial WLAN (iWLAN) system [50]. This protocol can only achieve low throughput since the protocol employs Time Division Multiple Access (TDMA) scheme and support single user transmission. Another Wi-Fi implementations for industrial applications is the iWLAN by Siemens [51] with special additional functions to meet the specific demands of Wi-Fi in industry. These are particularly useful for applications in automation, for example in automotive manufacturing, transportation and logistics, and the oil and gas industry. The penetration of mobile communication in industrial automation has so far been

low and mainly focused on remote service application and alert systems [52]. More recently, the suitability of cellular communication for industry automation has been highlighted and adopted for practical implementations.

2.5.2 Wi-Fi for V2X

The Intelligent Transportation Society of America (ITSA) has proposed Dedicated Short Range Communication (DSRC) wireless technology for vehicular communication which is based on IEEE 802.11p. Wireless Access for Vehicular Environment (WAVE) is an architecture proposed by the IEEE that defines the two communication modes Vehicle-to-Vehicle (V2V) and Vehicle-to-Anything (V2X). WAVE is a combination of IEEE802.11p and four other standards (1609.x). The IEEE 802.11p deals with Physical layer (PHY)/Medium Access layer (MAC) and IEEE 1609.X considers upper layers, as shown in Figure 2.6. The standard 1609.1 defines the resource manager service which enables remote applications to communicate with On-Board Unit (OBU) through Road Side Unit (RSU). The standard 1609.2 defines secure message formats and the processing of those secure messages in the WAVE system. In 1609.X standards family, 1609.3 define network and transport layers and 1609.4 specify the multi-channel operation. In the multi-channel operation, a WAVE system uses one common Control Channel (CCH) and several Service Channels (SCHs).

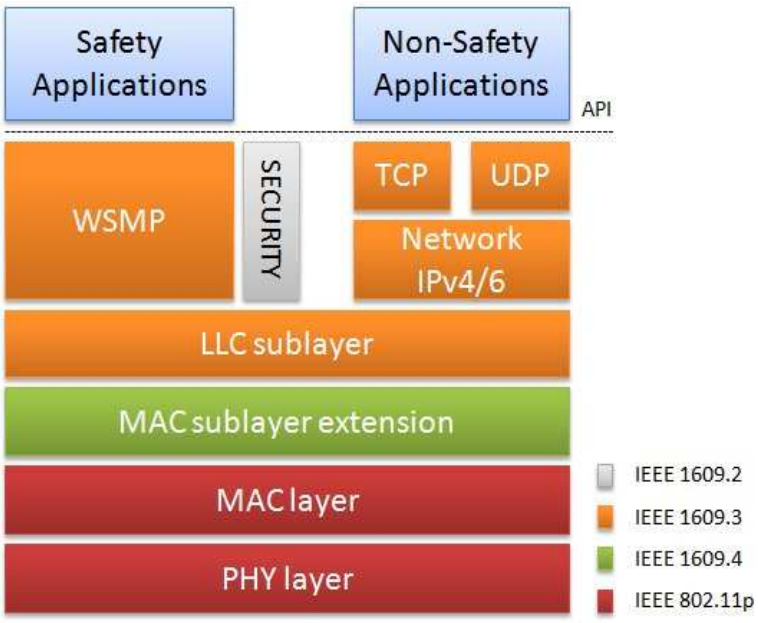

Figure 2.6: The IEEE802.11p protocol stack, according to [54].

OBUs and RSUs use 10 MHz bandwidth and operate in 5.9 GHz band. Supported data transmission rate ranges from 3 to 27 Mbps depending on the modulation scheme [55]. Using DSRC, vehicles equipped with the OBUs can communicate with each other and the RSUs for V2V or Vehicle-to-RSU (V2R) communication. V2V communication is critically important as

many safety applications rely on the safety messages broadcast among surrounding vehicles. Although 802.11p-based systems have been considered for a long time for V2X communication, cellular networks are nowadays being preferred over DSRC/WAVE due to the advantages offered such as already available infrastructure and wide range [56].

It is obvious from the above description of Wi-Fi that it supports high data rates, however, with a shorter communication range (<1000 m). In the use cases where a communication range of up to a few kilometers is required, for instance ITS, other alternatives such as cellular communication have many advantages over the use of Wi-Fi.

2.6 Cellular Networks for Machine Type Communication

The narrowband cellular technologies that are dedicated to MTC, i.e. LTE-M (officially known as CAT-M1) with its 1.4 MHz bandwidth and NB-IoT (officially known as Cat-NB1) with its 0.2 MHz bandwidth, also offer a nice alternative to the above mentioned LPWAN and Wi-Fi technologies. In fact, cellular networks provide more attractive options for low power and wide area networks due to the already existing infrastructure dedicated to much larger communication ranges. In this section, a detailed overview of LTE network, its protocol stack, and narrowband cellular network is presented. The high-level network architecture of LTE is comprised of the following three main components:

- User Equipment (UE),
- Evolved UMTS Terrestrial Radio Access Network (E-UTRAN),
- Evolved Packet Core (EPC).

Figure 2.7 illustrates the LTE network architecture. Its main components, that are UE, eUTRAN, and EPC as wells as its protocol stack are described in detail in the following subsections.

2.6.1 User Equipment

A UE in LTE can be seen as an end-user device, that connects/registers with the network in order to communicate. To this end it makes a network connection request. It can be a hand-held mobile phone, a laptop computer, or a stationary device e.g. installed on a machine. A Universal Subscriber Identity Module (USIM), which stores user-related information, is used normally inside a UE. The UE connects to the eNB in Evolved UMTS Terrestrial Radio Access Network (E-UTRAN). 3GPP has standardized different UE categories in its series of releases. Table 2.1 presents these categories from Release 8 to Release 14 for machine type

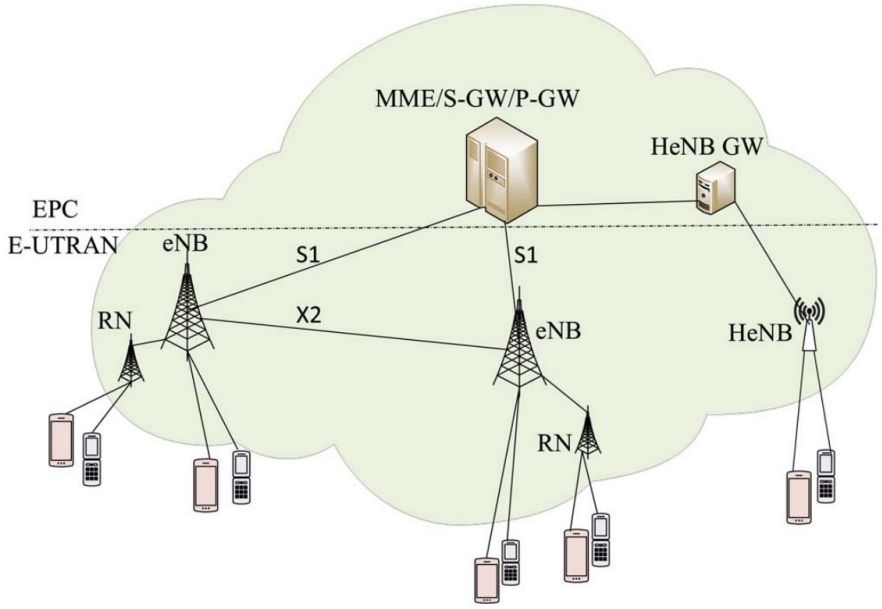

Figure 2.7: LTE E-UTRAN architecture according to [60]. Entities in the core network are connected with eNBs by S1 interface, whereas eNBs are connected to each other via X2 interfaces. A HeNB is a term used for femtocell. There are also RN in access network that forward the data to core network through a normal eNB.

communication [57–59]. There were no UE categories included in Release 9 to Release 11. Release 8 and 12 each included a UE category for MTC with relatively higher bandwidth (20 MHz). In Release 13, two UE categories named as LTE-M and NB-IoT were included in the LTE standards. NB-IoT is an extension of LTE networks specifically designed for long-range and low-energy use cases. These UE categories were included in LTE specifications to support the narrowband cellular communication, which provides better coverage, longer battery life and lower manufacturing cost for devices. Due to these benefits, LTE-M and NB-IoT are considered good candidates for MTC with cellular networks.

2.6.2 Evolved UMTS Terrestrial Radio Access Network

The architecture of the E-UTRAN for LTE is shown in Figure 2.7. The eNB is responsible for wireless communication with the UEs. The X2 interface is used between eNBs for message exchange. The wireless communication between UE and eNB is carried out in E-UTRAN, which includes synchronization signals, channel access procedures, scheduling messages, and data transmissions. The hand-overs are also managed by eNBs within E-UTRAN. The network radio resources are managed by eNB and allocated accordingly to the UEs that need to make uplink or downlink transmissions.

Table 2.1: 3GPP UE categories for MTC from Rel. 8 to Rel. 14 [57–59].

Release	Year	UE category	BW per channel	Downlink data rate per deice	Uplink datarate per device	Duplex mode	Max tx power
8	2012	Cat-1	20 MHz	10 Mbps	5 Mbps	Full duplex	23 dBm
12	2015	Cat-0	20 MHz	1 Mbps	1 Mbps	Half duplex (optional)	23 dBm
13	2016	Cat-M1 (LTE-M)	1.4 MHz	1 Mbps	1 Mbps	Half duplex (optional)	20 dBm
13	2016	Cat-NB1 (NB-IoT)	200 kHz	250 kbps	25 kbps	Half duplex	23 dBm
14	2018	Cat-NB2	200 kHz	85 kbps	150 kbps	Half duplex	14/20 dBm

*BW: Bandwidth, *NB: Narrowband

2.6.3 Evolved Packet Core

The EPC is responsible for the overall control of mobile devices and establishment of Internet Protocol (IP) packet flows. The EPC is a flat full-IP-based core network that can be accessed through 3GPP radio access (e.g., WCDMA, HSPA, and LTE/LTE-A) and non-3GPP radio access (e.g., WiMAX and WLAN), to efficiently access various services. The access flexibility to the EPC is attractive for operators since it enables them to modernize their core data networks to support a wide variety of access types using a common core network. The following text describes the main components of the EPC (see Figure 2.7) along with their functionalities.

- Mobility Management Entity (MME): The MME is a key control plane element for the LTE access network. It is responsible for managing security functions (authentication, authorization, and Non-Access Stratum (NAS) signaling), roaming, handover, and handling idle mode of the user equipment. It is also involved in choosing the Serving Gateway (S-GW) and Packet Data Network Gateway (P-GW) for a UE at an initial attach.

- Serving Gateway: The S-GW resides in the user plane, where it routes and forwards packets to and from the eNBs and P-GW. The S-GW is connected to the eNB through S1-U interface and to the P-GW through S5 interface. Each UE is associated to a unique S-GW, which will be hosting several functions.

- Packet Data Network Gateway: The P-GW provides connectivity from the UE to a Packet Data Network (PDN) by assigning an IP address from the PDN to the UE/M2M

device. Moreover, P-GW provides security connection between UEs/M2M devices by using Internet Protocol Security (IPSec) tunnels between UEs/M2M devices connected from an untrusted non-3GPP access network with the EPC.

A LTE system is considered as flat since from a user-plane point of view there are only the eNBs. This leads to a reduced complexity compared to previous architectures.

Another concept, which is recently developed in the context of industry automation is a private cellular network installed inside the business premises such as a factory hall or a large sized warehouse. Such private networks are named Campus Networks. They are tailored to the individual needs of users to meet future requirements of Industry 4.0. A campus network consists of a local radio access and core network. The private core network does not have the similar entities and functionalities as a typical core network. However, the idea of bringing the core network inside the base station is to reduce the deployment cost by eliminating multiple entities. Furthermore, it also reduces the additional delays that occur in core network. An illustration of a campus network is presented in Figure 2.8. The private campus network consists of a private base station, local server/core installed inside the base station, and an optional edge cloud. The campus network brings many benefits such as, reduced core network delay, improved security, and customizable network configurations.

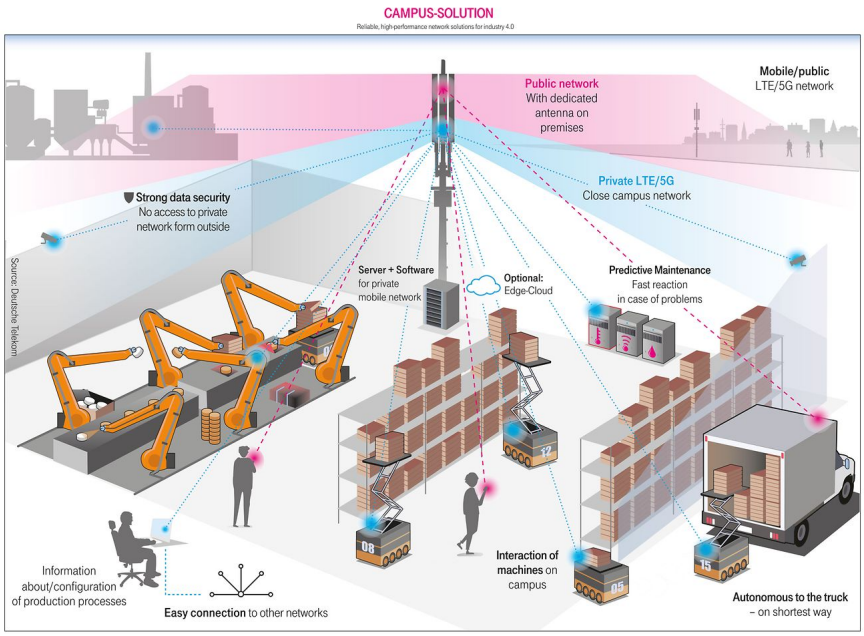

Figure 2.8: Illustration of campus network where a private base station including local core is installed inside the factory premises tailored to meet the individual requirements of users, according to [67].

2.6.4 LTE Protocol Stack

As mentioned earlier in section 2.6.2, the eNB provides the E-UTRAN with the control- and user-plane termination protocols. Figure 2.9 gives a graphical overview of both protocol stacks. In the both, control- and user-plane, the protocols include Packet Data Convergence Protocol (PDCP), Radio Link Control (RLC), MAC, and PHY layer protocols. The control-plane stack additionally includes the Radio Resource Control (RRC) protocol. The main functionalities carried out in each layer are summarized in the following [60].

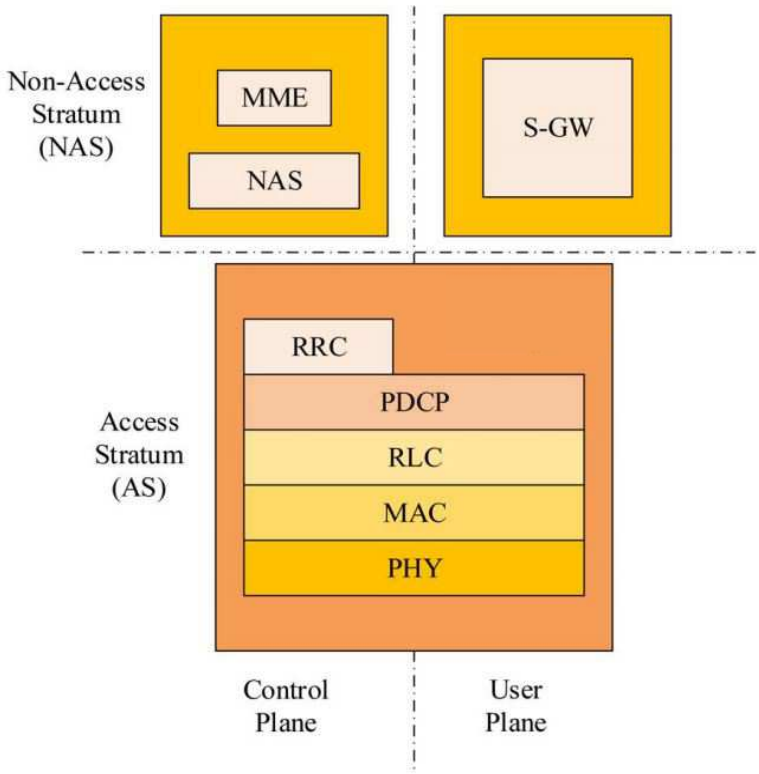

Figure 2.9: LTE control- and user-plane protocol stacks according to [60].

2.6.4.1 Non Access Stratum (NAS)

In the network structure, the NAS is the highest stratum of the control-plane between a UE and the core network at the radio interface. This layer is used to support the continuous connection of the UE as it moves, and also to manage the establishment of communication sessions to maintain IP connectivity between the UE and the P-GW. The P-GW connects the cellular network core to the Internet. Furthermore, the NAS is also a protocol for messages passed between the UE and the core network as can be seen in Figure 2.9. The work in this thesis does not concern the NAS, and only resides in the Access Stratum (AS). NAS messages include update or attach messages, authentication messages, service requests, and so

forth. In addition, the NAS control protocol performs bearer context activation/deactivation, registration, and location registration management.

2.6.4.2 Radio Resource Control

In the access stratum, which is used for the wireless communication between the eNB and UEs, the RRC protocol layer handles the control-plane signaling between the UE and eNB. The main services and functions of the RRC sublayer include broadcast of system information related to the NAS and Access Stratum (AS). Furthermore, establishment, modification, and release of RRC connections are performed in this protocol layer. Initial security activation (i.e., initial configuration of AS integrity protection and AS ciphering), RRC connection mobility including intra-frequency and inter-frequency hand-overs, and specification of RRC context information are the other important tasks of the RRC sublayer. Moreover, this sublayer performs QoS control functions, UE measurement configuration and reporting. In addition, the RRC transfers dedicated NAS information and non-3GPP dedicated information.

2.6.4.3 Packet Data Convergence Protocol

The Packet Data Convergence Protocol (PDCP) layer, on both eNB and UE side, performs IP header compression and decompression using Robust Header Compression (ROHC) protocol at the transmitting and receiving entities, respectively. Furthermore, the PDCP transfers user plane or RRC data, and this function is used for conveying data among users of PDCP services. Maintenance of PDCP sequence numbers for radio bearers and in-sequence delivery of upper layer Packet Data Unit (PDU) are other functions of the PDCP layer. In addition, the duplicate detection of lower layer Session Data Unit (SDU), ciphering and deciphering of user-plane data and control-plane data, and integrity protection of control-plane data are performed in this layer.

2.6.4.4 Radio Link Control

The RLC protocol layer exists in the UE and the eNB. It is part of LTE air interface control and user planes. This layer transfers upper layer PDU and performs error correction through Automatic Repeat request (ARQ). Moreover, the RLC protocol layer is used for concatenation, segmentation, and reassembly of RLC SDUs. In addition, re-segmentation and reordering of RLC data PDUs, RLC re-establishment, and error detection and recovery are the other functions of this protocol layer.

2.6.4.5 Medium Access Control

The MAC protocol is responsible for regulating the access to the shared medium. Furthermore, the MAC protocol has a direct bearing on the reliability and efficiency of network transmissions. This layer is in charge of multiplexing/de-multiplexing of RLC PDUs, scheduling information reporting, error correction through Hybrid ARQ (HARQ), logical channel prioritization, and transporting format selection.

2.6.5 Narrowband Cellular Network: LTE-M and NB-IoT

LTE networks will remain in place for the foreseeable future [61]. LTE was designed for high data-rate broadband services and therefore is not optimized for MTC data transfers, which are typically low data-rate, which is the rationale of a smaller bandwidth. MTC features were added to LTE in release 12 [61] to extend LTE services to machines for IoT communications. NB-IoT and LTE-M are two new technologies developed for cellular IoT applications.

LTE-M is the abbreviation for LTE Cat-M1 or LTE category M1. This technology is for IoT devices to connect directly to a cellular network, without a gateway. The bandwidth of LTE-M is 1.4 MHz with peak data rate of 1 Mbps. Both technologies (i.e. LTE-M and NB-IoT) are for low bandwidth cellular communications that connect to the Internet and transmit small amounts of data, with lower costs (both hardware and subscription) and long battery life. They are expected to connect hundreds of millions of things to the Internet in the next few years and there are some clear benefits that make these two technologies essential for the future of IoT communications.

NB-IoT is developed to enable efficient communication and long battery life for mass distributed devices and uses the already established mobile networks to connect these things. There are multiple features that can make NB-IoT lead the IoT communications market in a short future, the most important among them are highlighted below.

- Low cost: The NB-IoT modules have lower cost than the modules for other cellular communication technologies in the market (like, 3G, 4G, GPRS etc.) and also LTE-M. This cost is currently around 10 USD, and is expected to be between 5 and 7 in a few years [62].
- More connections per cell: NB-IoT devices use 180 kHz bandwidth so it is estimated the network supports more than 100,000 connections per cell for typical use cases with low-activity nodes.
- Excellent penetration in indoors and underground.

A summary for the comparison of network characteristics for LPWAN, Wi-Fi, LTE-M and NB-IoT is given in Table 2.2. Unlike the LPWAN technologies, cellular communication-based UE categories (i.e. NB-IoT and LTE-M) use licensed spectrum. The data rate provided by LTE-M is the highest among all compared technologies because it uses a comparatively larger bandwidth. On the other hand, NB-IoT offers the highest range. Cellular networks clearly have an advantage over the LPWAN technologies in terms of data rate, range and existing infrastructure.

Table 2.2: Theoretical comparison of various LPWAN technologies, Wi-Fi, and cellular network based low-power long-range UE categories, as described in [46].

Parameters	Wi-Fi	LoRaWAN	MIOTY	SIGFOX	LTE-M	NB-IoT
Range	Up to 1000 m	< 14 km	< 15 km	< 17 km	< 15 km	< 22 km
Frequency Spectrum	Unlicensed	Unlicensed	Unlicensed	Unlicensed	Licensed	Licensed
Bandwidth	20 MHz	125 kHz	2 kHz	100 Hz	1.4 MHz	200 kHz
Data Rate	54 Mbps	50 kbps	400 bps	100 bps	1 Mbps	200 kbps

To cover the wide range of requirements for the use cases considered in this thesis and discussed in section 1.5, even if other technologies such as Wi-Fi and LPWAN show potential in some aspects, only cellular networks are most suitable for addressing them. Therefore, in the following sections and chapters, the discussion is only limited to narrowband cellular network and its latencies. Before presenting the technical analysis of the proposed approaches in this thesis, the choice is also confirmed by the bibliographic analysis. In the following section, the recent developments in 5G cellular networks along with expected trends in 6G are discussed.

2.7 Current Developments in 5G and 6G

Since 3G, the cellular industry has been focused on increasing the data rate. This can be observed in major 5G features such as the Ultra-Dense Network (UDN), the Massive Multiple-Input Multiple-Output (MIMO), the system carrier bandwidth, and the beam forming techniques. As a result of these techniques, the spectral efficiency of 5G has more than tripled [63]. The data and the peak rate for 5G have increased more than 10-fold while the radio delay has decreased considerably.

According to Qi et al. [63], no new candidate technologies have been identified in 5G for the mMTC application that could surpass the existing 4G narrowband LTE to meet the key requirements for better coverage, cost sensitivity, and battery longevity. As a result, 3GPP has opted to improve the narrowband LTE system as a solution to delivering the mMTC application in 5G. Consequently, the three application pillars of 5G, i.e. eMBB, mMTC, and URLLC (see Figure 1.4), delivered by one 5G standard, actually require that the LTE systems also remain deployed and in use even with some further improvements to support mMTC and URLLC use cases. Therefore, improving the existing 4G LTE standards for supporting mMTC and URLLC is also essential. Moreover the work done in thesis on 4G LTE networks will also serve as part of 5G or even 6G cellular communication for MTC.

With the completion of 3GPP Release 15 of the 5G standard in early 2018, 2019 has become the first year for the commercialization of 5G. While the challenges faced by 5G are being resolved, research on 6G has already started. Currently, 6G and its possible drivers are not yet defined, given that the 6G performance requirements are not even available yet. Some of the expected trends for 6G as proposed by Qi et al. [63] are as following.

- 6G will continue to move to higher frequencies with wider system bandwidth,
- Massive MIMO will remain as a key technology,
- The data rate and spectral efficiency will continue to be the focus,
- 6G will continue to demand progress in the chipset density,
- Grant-free transmissions could be more prominent,
- mMTC is more likely to take shape,
- 6G will transform a transmission network into a computing network.

At this stage, it is too early to speculate what key technologies will mark 6G, since it is not known what the drivers for 6G might even be. However, the next generation system typically does not emerge from a vacuum. By examining the industrial and technological trends from previous generations, directions and trajectories associated with each new generation can be discovered.

2.8 Conclusion

To support machine type communication, the wireless communication technologies are expected to enable longer range, lower latency, and reduced energy consumption. Some of the available long range wireless technologies are discussed in this chapter. It has been observed that cellular communication-based technologies have an advantage over LPWAN and Wi-Fi mainly due to the available infrastructure, wide range, and comparatively lower latency. Furthermore, cellular network is a unique technology, which is agile enough to address different use cases. Therefore, in this thesis, cellular communication has been selected as a potential wireless technology to support URLLC use cases.

Mobile cellular networks have been evolving in the last three decades based on different application requirements. LTE networks have been deployed and being used for quite some time and are being evolved towards fifth generation of cellular networks. It is anticipated that improvements in LTE will also be a part of 5G networks to reduce deployment costs [64]. The evolution towards 5G is driven by the novel IoT use cases where for some of the applications, very low latency is required and the current LTE networks do not fulfill them. In the following chapters, a comprehensive analysis of LTE latency and its components is presented along with the latency reductions techniques and their evaluations.

3

Latency Analysis of 4G Cellular Networks

3.1 Résumé

De nombreux éléments contribuent à la latence dans un réseau LTE. Ces éléments ont un impact direct sur les performances du système et rendent difficile la satisfaction des exigences en matière de latence dans des d'utilisation critiques. La latence de la liaison montante dans un réseau LTE (Uplink Latency/UL) se compose de deux parties : la latence du plan de contrôle et la latence du plan utilisateur. Dans la latence du plan de contrôle, l'opération d'accès aléatoire par contention (Contention Based Random Access/RA) est basée sur un accès de type ALOHA, ce qui implique qu'en cas de demandes d'accès simultanées provenant d'un nombre important de dispositifs, les performances du réseau peuvent se dégrader et le fonctionnement de certaines applications critiques en termes de latence peut être compromis. Par exemple, en cas de collision de véhicules sur une route, un grand nombre de véhicules essaieraient d'accéder au réseau afin de signaler cet accident. Dans un tel scénario, les dispositifs qui émettent simultanément peuvent connaître une latence plus importante en raison des collisions sur les préambules. Une telle application nécessite un transfert de données sans interruption de l'ordre de 10 ms, alors que le réseau LTE actuel pourrait ne pas être en mesure de fournir une latence aussi faible. La latence du plan utilisateur est elle aussi constituée de plusieurs composants. La latence minimale de la liaison montante de ce plan en LTE est supérieure à 10 ms. Ce qui signifie que LTE, dans sa forme/norme actuelle, ne peut pas prendre en charge l'applications URLLC.

Dans ce chapitre, une analyse approfondie de la latence du réseau LTE 4G est présentée et la nécessité de l'améliorer pour les réseaux 5G pour des applications URLLC est étudiée. L'analyse présentée comprend à la fois les latences du plan de contrôle et du plan utilisateur. La latence du plan de contrôle inclut le délai de la procédure d'accès aléatoire (Random Access/RA). Dans ce chapitre elle est évaluée pour des équipements utilisateurs à bande étroite (Narrowband User Equipment/UE). L'analyse de la latence du plan utilisateur présentée dans ce chapitre consiste en des calculs théoriques et une évaluation par simulation pour chaque facteur pertinent de la latence de la liaison montante, c'est-à-dire de l'UE à la station de base du nœud évolué (eNB). La latence de la liaison descendante en LTE est inférieure à 10 ms et n'est donc pas abordée dans cette thèse. Ce chapitre évalue donc la latence des réseaux LTE et pose des questions concernant la réduction de la latence dans LTE.

3.2 Introduction

There are multiple components that contribute to the latency in a LTE network. These components have a direct impact on the performance of the communication system and make it a challenge to meet the requirements of latency-critical use cases. In this chapter, an in-depth analysis of E-UTRAN latency in 4G LTE network is presented and the need for improvements towards 5G networks for URLLC applications is discussed[1]. The presented analysis includes both the control-plane and user-plane latencies. The control-plane latency includes the Random Access (RA) procedure delay, which is evaluated for narrowband UEs[2]. The RA procedure is used by the UEs to connect to the eNB. The user-plane latency analysis consists of theoretical calculations and simulation evaluation for each factor relevant to uplink latency i.e. from UE to eNB. The downlink latency in LTE stays below 10 ms as explained in [65]. In the downlink, eNB allocates the access of network resources to UEs, thus, reducing the access request related delays. Therefore, the downlink latency is not brought under discussion in this thesis. A detailed description of ns-3 simulator design is also presented. The following sections bring the material for a good understanding of the latency components.

[1] A version of the evaluation for LTE-M has been published as a part of the original conference paper [56].

- Author did the literature review, evaluated the model, prepared the figures and manuscript.
- A. Sikora, B. Hilt, and JP. Lauffenburger reviewed the manuscript and helped in revising the manuscript.

[2] The design of ns-3 random access procedure for NB-IoT was conceptualized by the author and implemented by H. Benaissa as part of his Masters thesis.

3.3 LTE User-Plane Uplink Latency: A Theoretical Study

LTE uplink latency depends on multiple factors as presented in Figure 3.1. The UE waits for the uplink resources to send a scheduling request. The transmission of a scheduling request also consumes time equal to one Transmission Time Interval (TTI). The TTI is the transmission duration on the radio link by either the UE or the eNB. Afterwards, eNB decodes the scheduling request and sends a scheduling grant. Finally, UE sends data in corresponding resources after receiving a scheduling grant and processing it. Table 3.1 shows the respective durations for all of these steps. The UE performs the random access procedure before sending data from the network. The eNB allocates uplink resources in response to a request sent by an UE. After completing the RA procedure, the uplink latency in the user-plane mainly depends on TTI, resource scheduling delay and the processing of data units in eNB and UE. The uplink user-plane latency d_{ul} can be formulated as:

$$d_{ul} = t_{sr} + 3 \times TTI + t_{sg} + 2.5 \times t_{pr} \tag{3.1}$$

where t_{sr} is the duration to scheduling request opportunity, t_{sg} is the delay to scheduled resources, and t_{pr} is the processing time. It is worth noting that the TTI and processing time play a major role in uplink latency. The following subsections explain each of the component in detail.

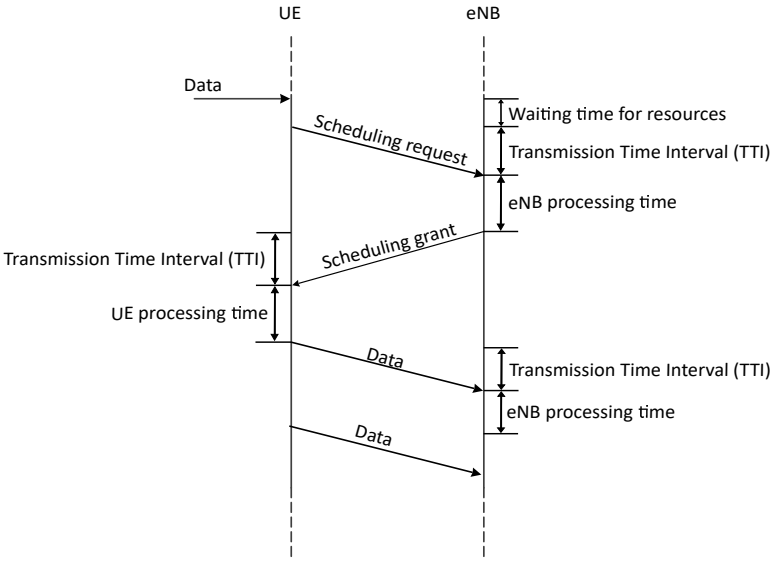

Figure 3.1: LTE user-plane uplink latency components, where transmission time interval and processing times are the major contributors.

Table 3.1: Uplink latency components for 1 ms transmission time interval [9], [66].

Component	Delay (ms)
Average delay to next scheduling request opportunity	0.5
SR transmission (TTI)	1
eNB processing delay	3
SG transmission (TTI)	1
UE processing delay	3
Average delay to scheduled resources	0.5
Uplink data transmission (TTI)	1
eNB processing delay	1.5
Total uplink latency	**11.5**

3.3.1 Uplink Latency Components

3.3.1.1 Grant Acquisition

In cellular networks, the network resources are managed by the eNB. The UEs that need to access uplink network resources, send a request to the eNB. This Scheduling Request (SR) is sent over the Physical Uplink Control Channel (PUCCH). In order to send a SR, the UE must wait for PUCCH SR-valid resource waiting time in Figure 3.1. The eNB first decodes the SR upon reception. This decoding processing time also contributes to the uplink latency. In response to the SR, the eNB sends a Scheduling Grant (SG) containing information about the scheduled resources. The resources are scheduled in the form of slots (7 OFDM symbols) and Physical Resource Block (PRB). The UE can start the transmission of its data over a Physical Uplink Shared Channel (PUSCH) after decoding the SG. Waiting for the PUCCH resources and transmission/reception of SR/SG causes a delay of roughly 6 ms (see first four rows in Table 3.1).

3.3.1.2 Transmission Time Interval

The transmission time interval refers to the transmission duration on the radio link. In LTE, a system frame has a length of 10 ms and consists of 10 sub-frames of 1 ms each. An illustration of transmission time interval is shown in Figure 3.2. The transmission of a request, grant, or data is executed in 1 ms sub-frames. Each transmission in LTE consists of 1 ms subframe, which is one of the major contributors in the uplink latency as evident from Table 3.1.

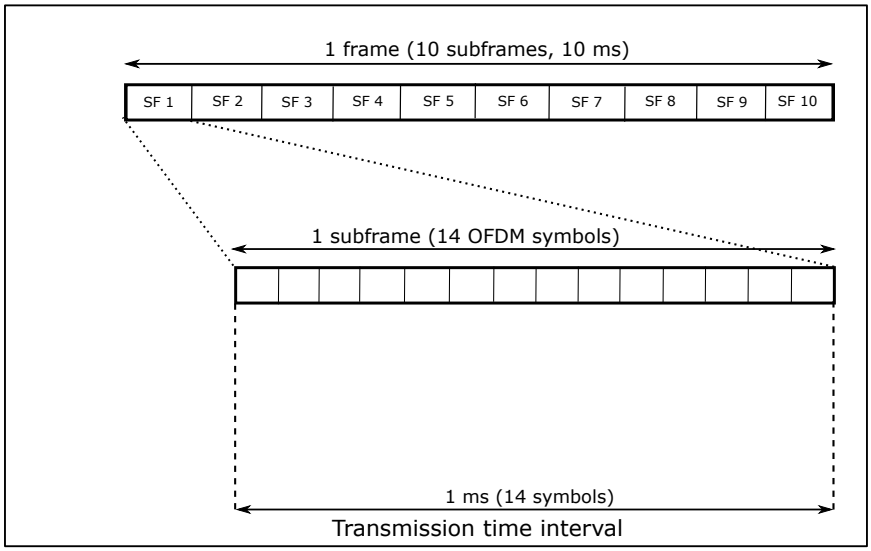

Figure 3.2: Illustration of transmission time interval. The subframe in LTE consists of 14 OFDM symbols and is equal to one TTI (i.e. 1 ms).

3.3.1.3 Processing and Core

The processing times, either from UE or eNB, is also a major contributor in LTE uplink latency. When a data unit is received at eNB or UE, it is processed in each layer starting from physical to application layer. The processing of each data unit also incurs an extra delay, which is normally equal to 3 times the TTI, as discussed in [9]. Both data and control messages need to be encoded/decoded in the network elements, which adds to the end-to-end delay. Congestion in the Core Network (CN) due to packet queues can insert additional delay in the system performance. However, CN delays can vary widely.

3.3.2 Towards a Realistic Simulation Tool for Experimental Validation

As discussed in section 1.6, among the available open-source simulators, ns-3 was selected for the simulations due to the fact that it provides the most comprehensive implementation of control- and user-plane of LTE protocol stack along with the core network. In ns-3, each layer of the LTE protocol stack is designed to take as input all the layer related parameters from the user. These parameters are defined in the simulation script along with the network node density, node positions, node mobility, and application on top of the LTE protocol stack. Ns-3 also contains a built-in pseudo-random number generator. By default, ns-3 simulations use a fixed seed. If there is any randomness in the simulation, each run of the program yields identical results unless the seed is changed. The random number generator module in ns-3 helps to overcome this issue by providing different seeds for the same simulation that is run multiple times. The LTE module also relies heavily on the pseudo-random number generator. The module generates output log traces in form of text files for each layer of LTE protocol

stack, which include a number of Key Performance Indicators (KPI). The latency is calculated in the PDCP layer and can be analyzed after the simulation completes from the respective output log trace. The following section further brings the material of a good understanding for ns-3 module design and simulations.

3.3.2.1 Ns-3 LTE Module

The legacy LTE module of ns-3, also called LENA [70], was initially developed by the Centre Tecnològic Telecomunicacions Catalunya (CTTC) in 2011 [68]. Over the period of last eight years, the LTE module has been enhanced by the research community and new features have been added. Figure 3.3 gives an overview of the LTE-EPC simulation model, which relies on two main components:

- LTE Model: this model contains the LTE Radio Protocol stack (RRC, PDCP, RLC, MAC, PHY). These entities are entirely resident within the UE and the eNB nodes as shown in Fig. 2.7 in Chapter 2.
- EPC Model: this model contains core network interfaces, protocols and entities [69]. These entities and protocols are resident within the S-GW, P-GW and MME nodes, and partially in eNB nodes.

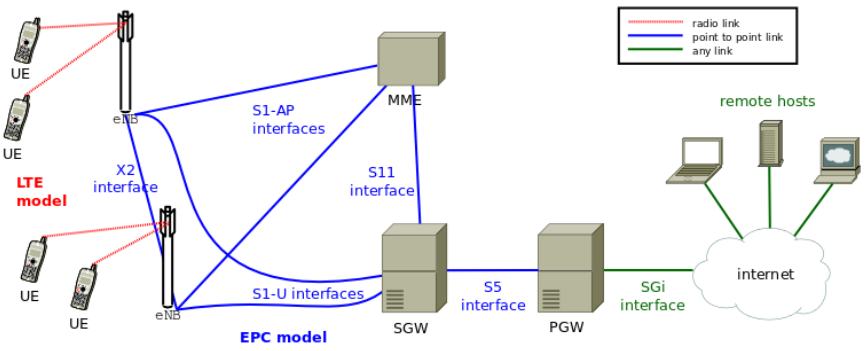

Figure 3.3: Overview of the ns-3 LTE-EPC simulation model according to [69]. The LTE network connects to the Internet through P-GW. The links colored in green are the ones that exist outside of the cellular network and could be any type of the link.

The LTE model was designed to support the evaluation of the various aspects of LTE systems:

- Radio Resource Management,
- QoS-aware Packet Scheduling,
- Inter-cell Interference Coordination,
- Dynamic Spectrum Access.

Ns-3 provides a generic node model for the simulations. Any communication protocol stack such as LTE, Wi-Fi, etc. can be installed during the simulation on top of the nodes. In the release at the beginning of the thesis, LTE module only supports wide-band UE categories (i.e. MBB). To enable support for narrowband UE, a NB-IoT UE class was implemented and evaluated in this thesis. For channel modeling purposes, the LTE module uses the spectrum channel interface provided by the spectrum module in ns-3. There are a number of different propagation models included in ns-3. The model specifically implemented to be used with the LTE module is the one provided by the Buildings module and named Hybrid Building Propagation Loss Model. It can also be used with other wireless technologies. The LTE module includes a trace-based fading model. The main characteristic of this model is the fact that the fading evaluation during simulation run-time is based on per-calculated traces. This is done to limit the computational complexity of the simulator.

The EPC model implements end-to-end IP connectivity in simulation on the LTE model. It supports the interconnection of multiple UEs to the Internet, via a radio access network of multiple eNBs connected to a single S-GW/P-GW node. The simulations only support IPv4 as PDN type. The EPC model is designed to simulate the end-to-end performance of a realistic use case, so it allows the use of any regular ns-3 application working on Transmission Control Protocol (TCP) or User Datagram Protocol (UDP). Figure 3.4 shows a representation of the end-to-end LTE-EPC data plane protocol stack as implemented in the simulator.

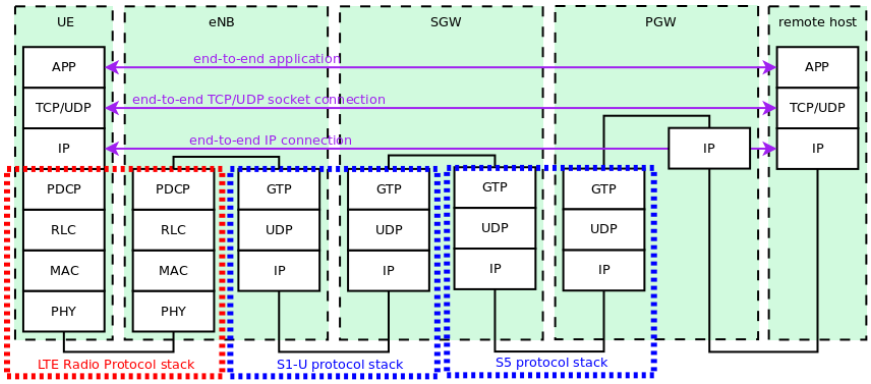

Figure 3.4: LTE-EPC data plane protocol stack implemented in ns-3 according to [69]. The IP module lies inside the P-GW as the connectivity to the Internet is maintained through P-GW in cellular network.

A simulation in ns-3 is run using a simulation script that defines the network topology, communication technology used, and other network related parameters. A generic block diagram of ns-3 simulation for LTE module is shown in Figure 3.5. The simulation scenario parameters are given to the simulation script which is then passed to the model inside ns-3. During the execution if the simulation, the log traces are generated by the model and shown to the user at the end of the simulation. These log traces help in analyzing the behavior of the model.

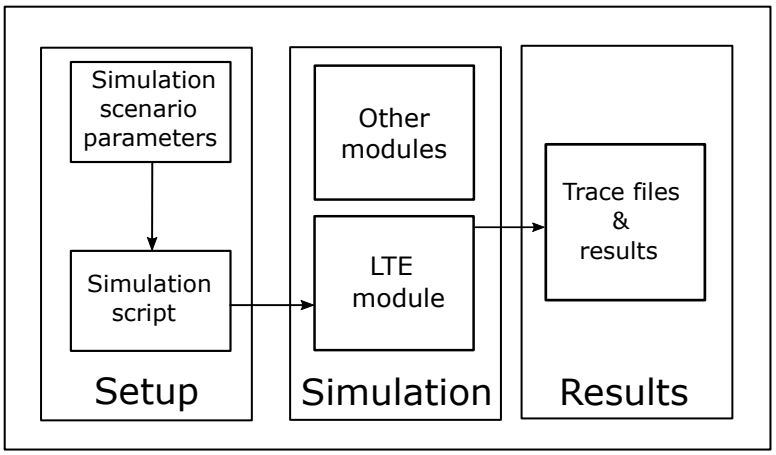

Figure 3.5: A block diagram of simulation setup in ns-3 for LTE module.

In fact any simulation of LTE follows a similar flow as the one shown in Figure 3.6. The first step is always to define the global network parameters. After that, the initialization of nodes and respective protocol stack on top of them is initiated. The initial node positions and the mobility model are also defined for the nodes. Towards the end of the simulation flow, as shown in Figure 3.6, an application is defined on top the protocol stack to start the data exchange between nodes and the network. At the end of the simulation, different log traces, that are generated during the simulation, are available as output log files for the analysis.

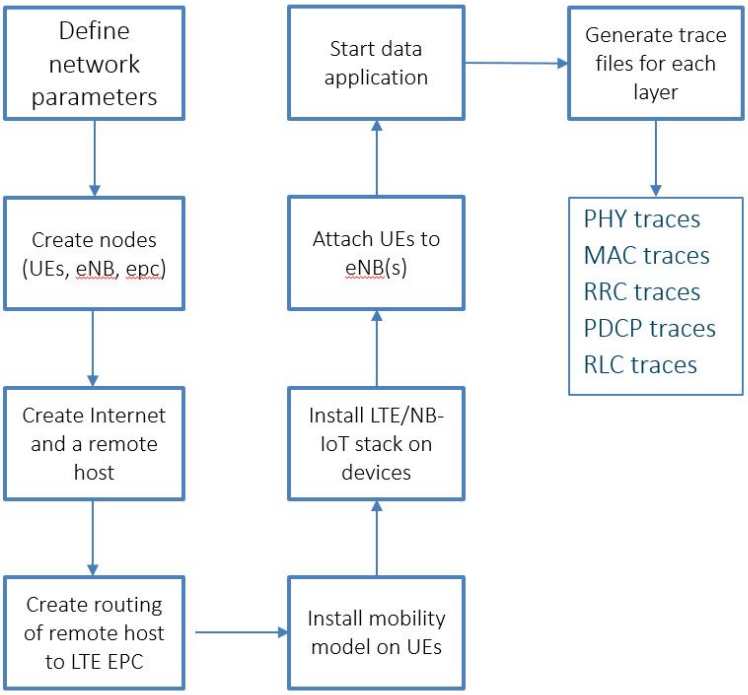

Figure 3.6: The flow of ns-3 simulation. Any simulation in ns-3 follows the similar flow in order to define all the parameters, network topology and data application.

To have a simulation tool that reproduces the data plan process as accurately as possible to determine the performance and limitations of the entire uplink data transmission process, the

following section presents the evaluation of user-plane latency of narrowband LTE networks through simulation. With the aim to investigate narrowband UE category for MTC, LTE-M with 1.4 MHz bandwidth is used for the simulations. The following subsections describe the overall simulations setup and evaluation of the obtained results.

3.3.2.2 Simulation Setup

As discussed earlier, the ns-3 simulator includes a well-developed LTE module called LENA. The LENA module does not support NB-IoT model. That is why, in this thesis, a NB-IoT module for ns-3 has been developed and validated. These features include NB-IoT UE class support, RRC and PDCP layers, and resource scheduling. The bandwidth is defined by the number of resource blocks and is used as 100, 6, and 1 for Cat-0, LTE-M, and NB-IoT respectively. A resource block is comprised of 12 frequency sub carriers of 15 kHz each. The user-plane latency is measured as the delay between the PDCP layers of the eNB and the UE. The UEs in the simulations are placed randomly in a circle around eNB with a radius of 2 km. All UEs during the simulation send data to eNB periodically after a predefined time interval. The parameters used in our simulations are shown in Table 3.2. The results shown in this section are means calculated from 10 independent simulation runs.

Table 3.2: Simulation parameters used in ns-3 simulations for the evaluation of control-plane and user-plane uplink latencies.

Parameter	Value
Simulator - version	ns - 3.26
Propagation loss model	Two-Ray Ground
eNB transmission power	43 dBm
UE transmission power	{20, 23} dBm
Uplink bandwidth	200 kHz, 1.4 MHz, 20 MHz
Number of resource blocks	1, 6, 100
Transmission time interval	1 ms
Resource scheduling	Dynamic
Simulation duration	30 s

3.3.2.3 Results

The goal of the simulation evaluation of user-plane latency is to validate the analytical calculations of uplink latency with the simulated one. As stated earlier, the uplink user-plane latency is measured as the time between the data is passed from the application to the LTE protocol stack on the UE until the time when the data is processed by the LTE protocol stack

in the eNB and available to be used by the application. Figure 3.7 presents the comparison of the LTE uplink latency for LTE-M (1.4 MHz) and Cat-0 (20 MHz) from simulations with an increasing number of devices (from 20 to 180). The uplink latency remains higher than 10 ms for all UE configurations. For LTE-M, the latency increases slowly up to 100 UEs and then increases abruptly. This sudden increase happens when the available channel capacity is lower than the total data traffic offered from UEs. Furthermore, the increase in latency after resource saturation is linear because the scheduling in our simulations is performed in a round robin process. Therefore, the impact of an increasing number of UEs includes a linear latency increase. The latency for Cat-0 UEs also follows the same linear increase after resource saturation.

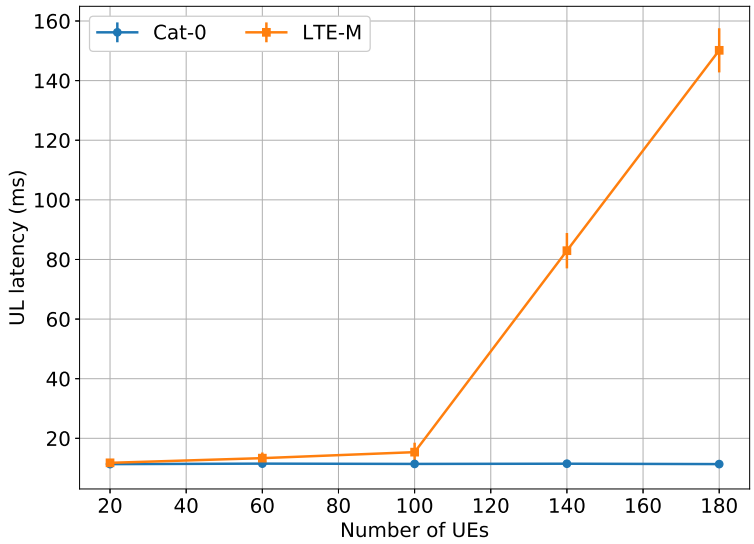

Figure 3.7: User-plane uplink latency comparison of LTE-M and Cat-0 in LTE network. For both categories, the minimum latency is above 10 ms not meeting URLLC latency requirements. The latency of LTE-M remains low as long as the offered load from devices is below the channel capacity (traffic from 100 devices in this case). Beyond this point, the latency increases significantly.

The performance of different scheduling algorithms has been already investigated by Dawaliby et al. [71]. The maximum uplink data rate defined for LTE-M is 1 Mbps. The total data traffic offered to the network depends on the number of devices in the network cell and the application data rate, which is used to send data over the network. In Fig. 3.7, each UE sends a data packet of 12 bytes every 100 ms, which results in a data rate of 9.6 kbps per UE. In the case of LTE-M, the total data rate sent to the network by 100 UEs is 960 kbps, which is less than the peak data rate of LTE-M [71]. The data rate produced by 140 UEs is 1.34 Mbps, which exceeds the network capacity and overloads its resources. LTE-M with 1.4 MHz bandwidth and 1 Mbps peak uplink data rate is clearly not suitable for high data

rate (>1 Mbps) applications; however, it provides a promising potential for low data rate applications that require very low latency due to lower cost and higher range.

Similarly, Figure 3.8 presents the comparison of uplink user-plane latency for NB-IoT and LTE-M. Since the available system capacity in terms of data rate is much lower in NB-IoT due to the narrow bandwidth of 200 kHz, the resource saturation starts earlier, where the total sent data traffic from the UEs is higher than the total system capacity of 20 kbps in uplink.

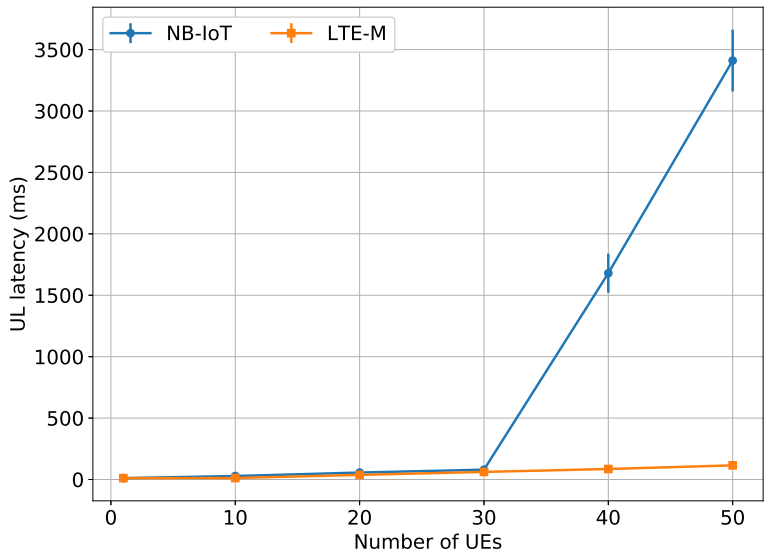

Figure 3.8: User-plane uplink latency comparison of NB-IoT and LTE-M in LTE network. The minimum latency for both categories is above 10 ms, which cannot fulfill URLLC latency requirements. The latency of NB-IoT remains low as long as the offered load to the network from devices is below the network capacity (traffic from 30 devices in this case). Beyond this point, the latency increases significantly.

3.4 LTE Control-Plane Uplink Latency

In the LTE standard, in order to connect to a network, a UE receives multiple synchronization signals from the eNB when it is powered on. These signals include Primary Synchronization Signal (PSS), Secondary Synchronization Signal (SSS), Master Information Block (MIB), and System Information Block (SIB). These signals include the information about the synchronization, frequency band and bandwidth, network parameters for a UE, and information about the channel access. After receiving the information about channel access, a UE can perform the RA procedure to get access to the network resources. In LTE, channel access is ALOHA-type access i.e. all contesting devices transmit in the first available opportunity. Due to the limited number of total network resources for channel access, the RA procedure brings a significant amount of latency in the system. This mainly results from similar random selections of RA preamble that lead to collisions. The following subsection describes each

component involved in the LTE RA procedure and sheds light on their potential issues that bring the higher amount of latency. These components are later evaluated in subsection 3.4.2.

3.4.1 Random Access (RA) Procedure

To send/receive user-plane data, a UE needs to access network resources. The eNB allocates resources to UEs after the completion of an access procedure. A UE performs this random access procedure whenever:

- New data is available to transmit but without existing uplink synchronization. A UE can only transmit on uplink resources if it is time-synchronized,
- Recovering from a link failure,
- Changing of state from *idle*[3] to *connected*[4],
- Performing a handover in case the UE needs to synchronize with a new eNB due to the mobility.

3.4.1.1 Types of RA Procedure

There are two different types of random access defined in LTE [56]:

1. *Contention-based:* the UE tries to connect to the network by randomly selecting a preamble[5] from a pool of predefined ones. Due to this random selection, collisions can occur in this type of access.
2. *Contention-free:* the eNB informs the UE, which preamble to use for access procedure in order to avoid collisions. This type of random access is mainly used for the handover procedure and therefore not considered here.

3.4.1.2 RA Characteristics

To access the network resources, a contention-based random access procedure is performed by a UE on a dedicated physical channel called Physical Random Access Channel (PRACH). The PRACH spans six PRB and its periodicity is defined by the PRACH Configuration Index parameter, which is transmitted by the eNB. The PRACH periodicity varies, depending on the network configurations, from a minimum of one slot in every two frames (i.e. 20 ms), also called index 0, to a maximum of one slot in every sub-frames (i.e. 1 ms). Figure 3.9 illustrates

[3]The radio is inactive in this state but the IP address is assigned. The UE is known to EPC but unknown to eNB.
[4]The radio is active in this state and UE is known in both EPC and eNB.
[5]Preamble is a sequence defined by Zadoff-Chu codes [72] by performing cyclic shift.

different PRACH Configuration Index options. The blue highlighted subframes are the ones where UEs can initiate the RA procedure.

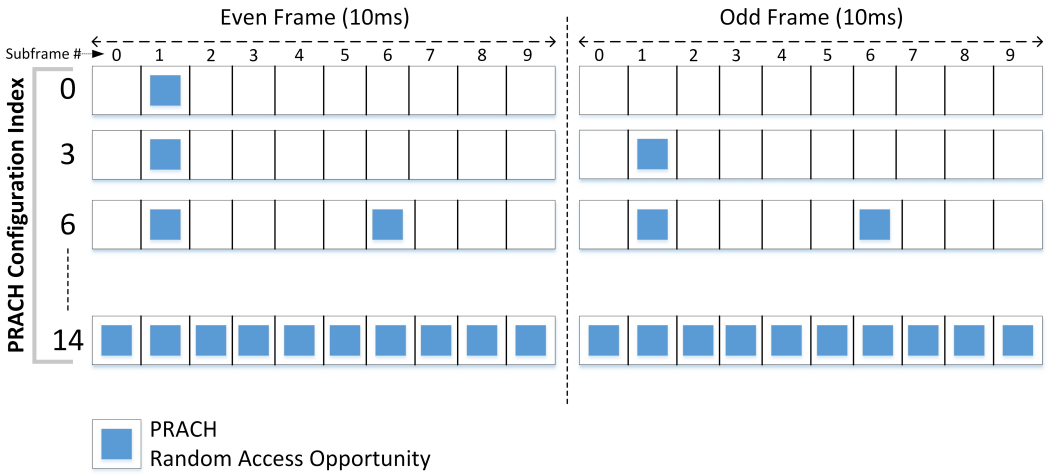

Figure 3.9: PRACH Configuration Index example. The minimum value of PRACH Configuration Index is 0, which corresponds to one RA opportunity every two frames (the upper most case), while 14 represents RA opportunity in every sub-frame (the lower case).

The length of a RA slot in time domain is defined by the format of the access request and varies, depending on the network configurations based on the channel quality, from 1 ms to 4 ms. Figure 3.10 illustrates the contention-based random access procedure, which consists of four messages exchanged between the UE and the eNB.

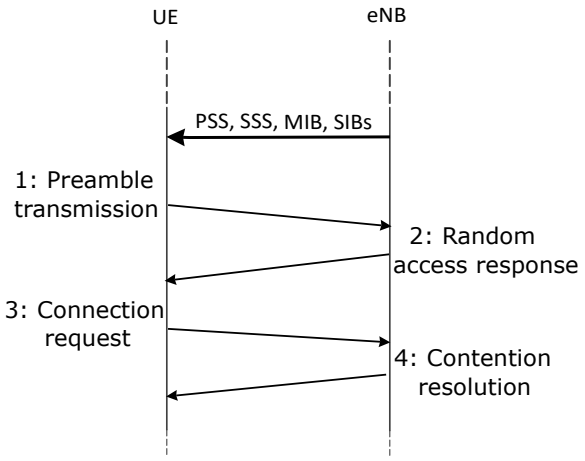

Figure 3.10: Contention-based Random Access message flow in LTE. 1) Preamble transmission from UE to eNB: an indication from UE to eNB for an attach request, 2) eNB responds with RAR message, which contains the scheduled resource for Message3, 3) UE sends a connection request, 4) eNB informs UE about completion of RRC connection.

3.4.1.3 RA Procedure Components

Contention-Based RA, 1st Message - Preamble Transmission This is the first message issued by the UE to the eNB (see Figure 3.10). Whenever a UE requires access to network resources, it selects the next available RA slot to send the access request that

includes a preamble. There are 64 pseudo-random preambles in total. The information about available preambles for contention-based RA is periodically disseminated by the eNB. The eNB reserves some of the preambles for the handover procedure and can detect different preambles transmitted in the same RA slot due to the orthogonality of these preambles. However, randomly chosen identical preambles, transmitted by two or more UEs in the same RA slot, result in a collision, since the eNB cannot decode the preamble in this case. Furthermore, a collision can go undetected if one of the colliding preambles is received with a higher SNR, as explained by Polese *et al.* [73]. After sending the preamble, the UE waits for a time window starting 3 subframes after the preamble transmission to receive the second message of the RA procedure from eNB. The eNB broadcasts the duration of the waiting window that is defined between 2 ms and 10 ms [74]. The preamble for each access request is randomly selected among those available for the contention-based random access.

Contention-Based RA, 2nd Message - Random Access Response After having successfully decoded the received preambles, the eNB computes a number called Random Access Radio Network Temporary Identifier (RA-RNTI) for each preamble. RA-RNTI is computed based on the RA slot used to send the preamble. As shown in Figure 3.10, the eNB then sends a Random Access Response (RAR) which includes a temporary identifier, referred to as Cell Radio Network Temporary Identifier (C-RNTI), detected preamble index, an uplink scheduling grant for UE to send the connection request, and a timing alignment for the UE to synchronize with the eNB.

The RAR is addressed to all UEs that send the preamble message in a specific RA slot. If the received RAR does not contain the preamble identifier (associated to the preamble sent by the UE), the UE performs a random back-off time. A collision can go undetected, if two UEs are at the same distance from the eNB and the eNB receives the same preamble constructively from both devices. In this case, eNB adds the preamble identifier and scheduled uplink resource for the third message in RAR and eventually a collision occurs in Message-3, as both UEs with undetected collision transmit at the same time.

Contention-Based RA, 3rd Message - Connection Request The UE transmits a connection request message, which includes C-RNTI and the reason for access request. With the transmission of a connection request, the UE initiates a contention resolution timer. An undetected preamble collision can lead to a collision again in this phase, as multiple UEs might transmit connection request simultaneously. In this case, eNB cannot send an acknowledgment for connection request and the UE retransmits connection request. Upon reaching maximum

number of allowed connection request retransmissions, the UE declares access request failure and starts the RA procedure again.

Contention-Based RA, 4th Message - Contention Resolution Upon receiving connection request, the eNB responds with a contention resolution message as shown in Figure 3.10. A UE will declare a failure to access request if it does not receive any contention resolution message, and, in such case, starts the RA procedure from the preamble transmission phase again. With each unsuccessful access attempt, the UE increments the preamble transmission counter. The network is declared unavailable when the counter reaches the maximum allowed preamble retransmissions.

3.4.1.4 Limitations of the RA Procedure

It is clear from the description above that the RA procedure in LTE is based on ALOHA-type access, i.e. all contesting devices transmit in the first available opportunity. This also means that with a higher device density in the network, more devices try to contest for the channel access and send their preamble. Such large amount of preamble requests leads to more collisions and an increased access delay. For MTC applications, the RA procedure becomes a bottleneck problem where hundreds of devices could potentially start access request simultaneously. Another issue with the RA procedure is the access latency, which results from performing contention-based RA. For URLLC applications, such delay becomes a hurdle for the network to fulfill the latency requirements.

3.4.2 Experimental Evaluation of the Random Access Procedure

This section presents the evaluation of the random access procedure for LTE-M and NB-IoT UE categories. The aim of this section is to evaluate the delay that UEs with limited bandwidth undergo while trying to access the LTE network. The current implementation of the LTE module in the standard version of ns-3 uses ideal random access procedures, i.e. only the first two messages of the RA procedure are modeled. Moreover, in this configuration, RACH is not subject to radio propagation as well. To mitigate this lack of realism, the work from Polese *et al.* [73] is used here only for the evaluation of LTE-M random access procedure. They implemented a *realistic RACH*[6] model for ns-3 LTE module. Their work focuses on wide-band (5 MHz) M2M communication. It is important to note that only LTE-M random access procedure is evaluated with ns3.24. All other evaluations and implementations in this these are carried out using ns3.26.

[6]https://github.com/signetlabdei/lena-plus

The NB-IoT UE category support is not available in the current release of ns-3 simulator. The NB-IoT features development is one of the goals of the work in this thesis, therefore, the RA procedure for NB-IoT devices is developed and evaluated in this section. For NB-IoT RA procedure, the implementation model of RA from ns-3.26 is used, which means there are only first two messages of RA modeled in the simulations. The development of NB-IoT module was started by University of Florence [75] in ns-3.26. Afterwards, the PHY layer was developed by a team in University of Antwerp [76], where they have mainly focused on power saving modes of NB-IoT. Many features of NB-IoT module including RA procedure, resource scheduling, UE category support and higher layers were developed as part of this thesis. Due to all the developments in ns-3.26, the RA model of the LTE module is followed, and therefore, the last two messages of NB-IoT RA is not modeled in the simulations.

3.4.2.1 Simulation Setup

Table 3.3 presents the network simulation parameters. For the evaluation, simulations were performed for 3GPP LTE-M and NB-IoT UE categories with a bandwidth of 1.4 MHz and 200 kHz respectively. After obtaining the system information transmitted by the eNB, the UEs start their RA procedure. The results shown in this section are means calculated from 10 independent simulation runs.

Table 3.3: Simulation parameters used in ns-3 for the evaluation of LTE-M and NB-IoT random access

Parameter	Value
Simulator version	3.24 for LTE-M, 3.26 for NB-IoT
Propagation loss model	Friis Propagation Model
eNB transmission power	43 dBm
UE transmission power	{20, 23} dBm
Uplink bandwidth	1.4 MHz, 200 kHz
Number of resource blocks	6, 1
PRACHConfigIndex	{0, 3}
Number of available preambles	{48, 54, 60, 64}
Packet interval time	100 ms
Number of eNBs	1
Simulation duration	30s

3.4.2.2 Impact of RA parameters on LTE-M RA Latency

Access Delay Figure 3.11 presents the average time to complete the RA procedure and the number of collisions depending on the number of UEs trying to access the network

simultaneously. The vertical lines at each point in the graph represent the standard deviation from independent simulation runs.

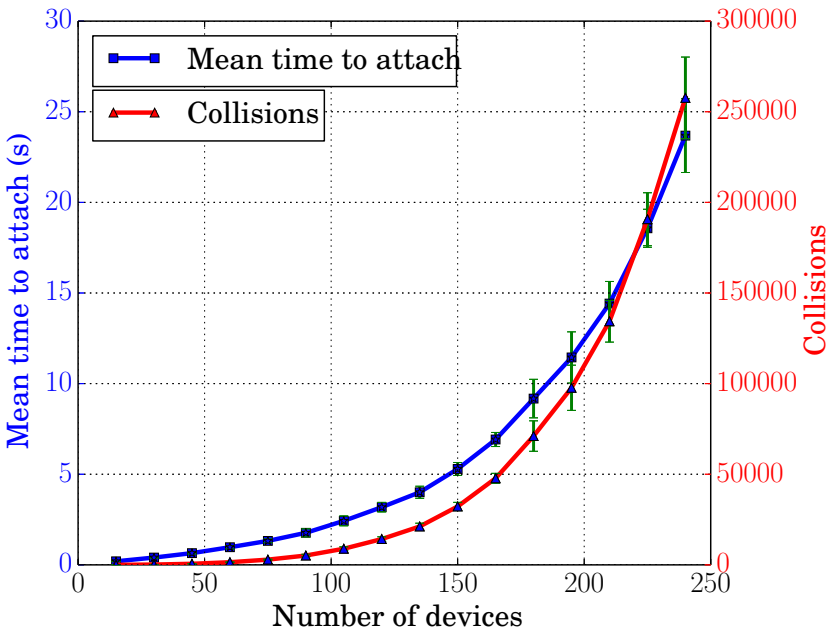

Figure 3.11: LTE-M: Mean time to complete the RA procedure (left y-axis), total number of collisions in the first preamble message (right y-axis); with increasing number of UEs simultaneously sending access requests.

It can be observed from these results that with the increase in number of devices that connect simultaneously to an eNB, the access delay has an exponential like increase. Such large number of simultaneous access requests are possible in MTC applications, such as, in case of collisions and cooperative driving, where a large number of vehicles/devices try to access the network to exchange information. The increase in preamble collisions also follows an exponential like growth as the access requests increase. This is due to the limited number of preambles available in the system that are selected randomly by the contesting UEs. It is also important to note that for more than 200 UEs, the number of collisions has a higher rate of increase than that of time to attach. The reason behind this higher rate of increase is that on some instances, the preamble collision ca go undetected at eNB resulting in another collision occurring at the transmission of message 3 as explained earlier.

It is worth noting that an increased system bandwidth improves the performance of the RA procedure. As presented in [73], with a bandwidth of 5 MHz (i.e. 25 PRBs), it takes around 1 second for 50% of the UEs to complete random access in case of 200 simultaneous access requests. The access delay has an inverse relation with available bandwidth (number of resource blocks) i.e., fewer the number of available resource blocks, the larger is the access delay. As compared to 5 MHz bandwidth, the duration to complete RA is 10 times higher in 1.4 MHz bandwidth. The approach in [73] is intended to evaluate massive number of arrivals for MTC;

however, using a bandwidth of 5 MHz for MTC might not be practical due to higher device cost, shorter range, and shorter battery life. Moreover, there are other standardized categories for MTC UEs by 3GPP, such as LTE-M with 1.4 MHz bandwidth, Narrowband-IoT [10] with 200 kHz bandwidth.

Number of Preambles Two series of simulations were carried out here for an in-depth analysis of LTE Cat-M RA. In the first part of these tests, the number of available preambles is varied while keeping all other RACH parameters fixed according to Table 3.3. Figure 3.12 presents the mean time taken by the UEs to complete the access procedure for different numbers of available preambles.

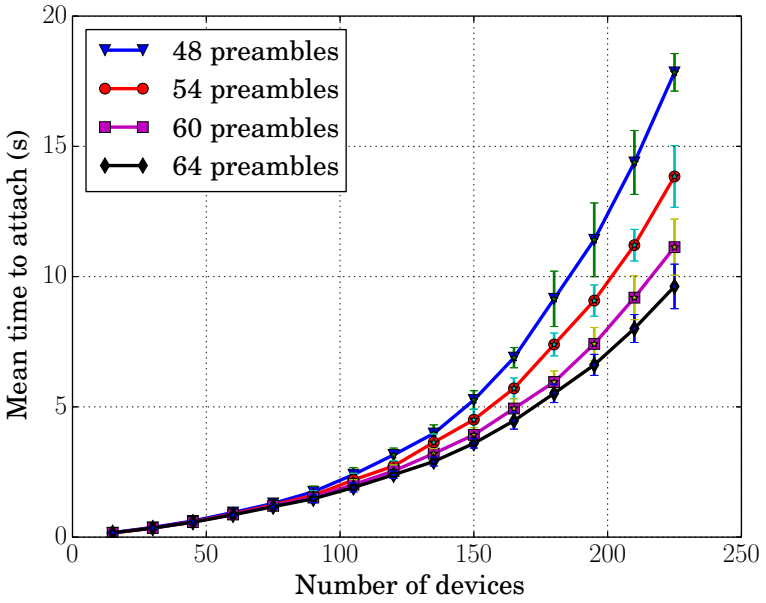

Figure 3.12: LTE-M: Mean time to complete the RA procedure with different numbers of available preambles.

Obviously, the best case, where all 64 preambles can be used by the devices performing contention-based random access, requires the least time, whatever is the number of devices. The difference in RA completion time remains below 2 sec for less than 100 devices. As the number of UEs increases, the difference also grows. For 225 simultaneous access requests, the mean access delay almost doubles between 48 and 64 preambles due to the fact that with 48 preambles, the number of collisions are higher than with the 64 preambles. However, many use cases of URLLC, such as V2X, include mobility of UEs with high speeds, which leads to handover of UEs between serving cells and utilizing a few preambles reserved for contention-free random access. Therefore, the best case i.e. 64 preambles might not be practical in those use cases because the network must have to reserve some preambles to be used for handover procedure based on contention free random access.

Figure 3.13 shows the distribution of UEs that completed RA procedure over time. The results show that for a larger number of simultaneous arrivals (i.e. 200 UEs) and with the lowest number of available preambles (i.e. 48 preambles), the time to attach for initial 50% of UEs is much higher than for the remaining half of UEs. This is explained by the fact that with larger number of UEs, more collisions occur, which results in a higher access delay. On the other hand, the more UEs succeed in attach procedure, the less the collisions of remaining unattached UEs.

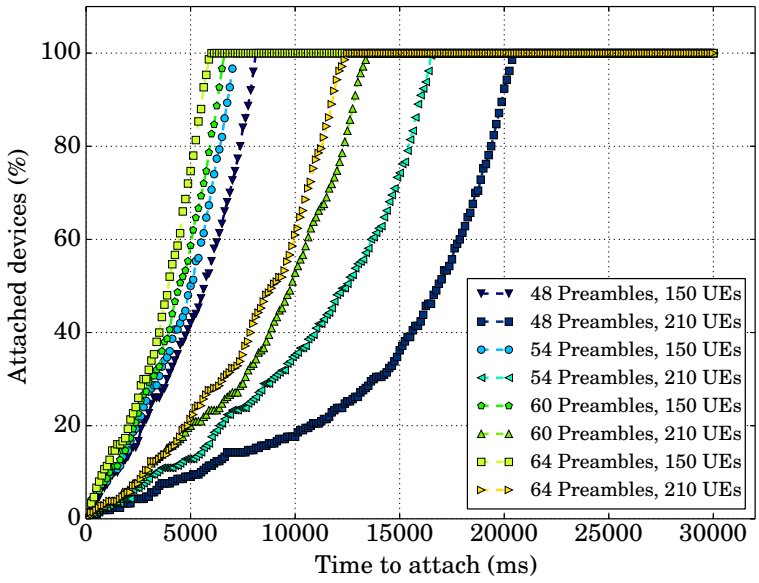

Figure 3.13: LTE-M: Percentage of attached devices over time with different number of available preambles.

PRACHConfigIndex In a second series of simulations, the effect of *PRACHConfigIndex* on the mean RA completion time is analyzed while keeping other parameters fixed. *PRACHConfigIndex* defines the time interval between two consecutive RA preamble transmission slots. As presented in Figure 3.9, *PRACHConfigIndex* 0 corresponds to one RA slot per two frames (i.e. 20 ms) while *PRACHConfigIndex* 3 represents one RA slot per frame (i.e. 10 ms). It is obvious from the result that for a larger number of UEs, the time to attach for *PRACHConfigIndex* 0 is three times more than the time to attach for *PRACHConfigIndex* 3 (see Figure 3.14).

Figure 3.15 shows the combined impact of number of preambles and *PRACHConfigIndex*. There is a considerable decrease in the mean RA completion time with more preambles and RA slots. The right choice of RA parameters can lead to an improved performance from the system. However, RA procedure limits the minimum achievable latency of the system. Before the start of uplink data transmission, the UEs are required to perform random access procedure in order to switch to connected state. They can only transmit data after RA procedure is complete

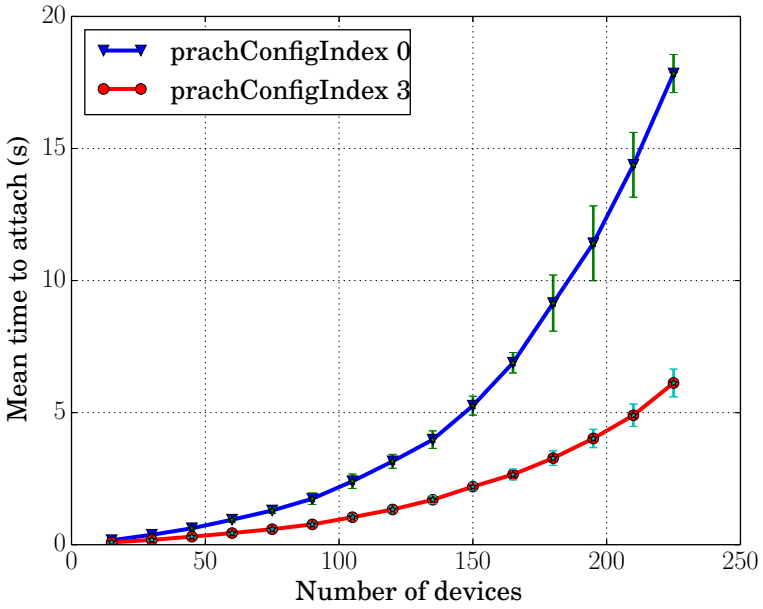

Figure 3.14: LTE-M: Mean time to complete the RA procedure with different numbers of RA slots per frame. *PRACHConfigIndex* 0 corresponds to one RA slot per two frames, whereas *PRACHConfigIndex* 3 equates to one RA slots per frame.

and the uplink resources are granted. Therefore, the total uplink latency including control and user-plane, is effected by the time it takes to complete the random access procedure.

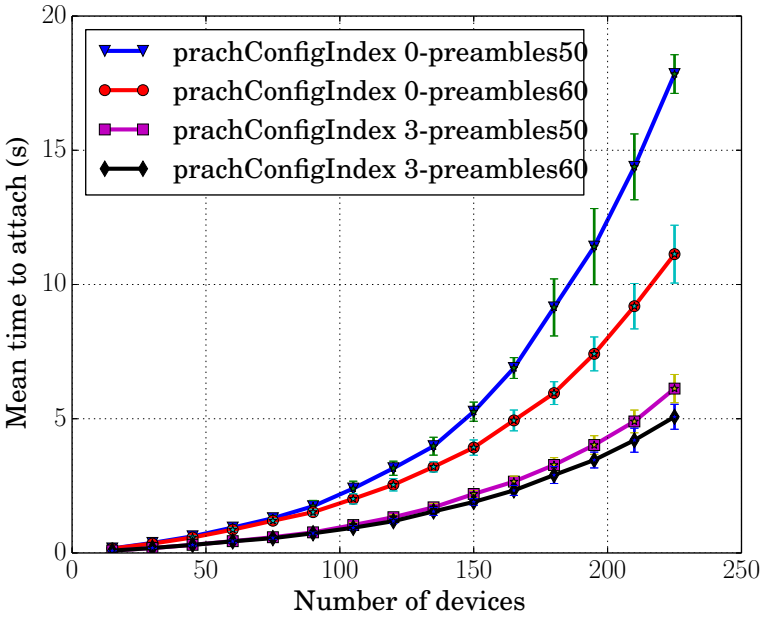

Figure 3.15: LTE-M: Mean time to complete the RA procedure with different number of RA slots per frame and number of preambles.

3.4.2.3 Impact of RA parameters on NB-IoT RA Latency

Access Delay Figure 3.16 shows the mean time to complete the RA procedure and the number of collisions with an increasing number of UEs trying to access the network simultaneously. These results show that as the number of devices increases, the access delay increases. The results of NB-IoT RA procedure exhibit exactly the same behavior as LTE-M RA. However, the total time to complete the RA procedure in NB-IoT is less than that in LTE-M. It is due to the fact that for LTE-M, all four messages of RA are implemented by extending the standard LTE module. On the other hand, in NB-IoT extension, only first two messages are implemented following the standard NE-3 LTE implementation.

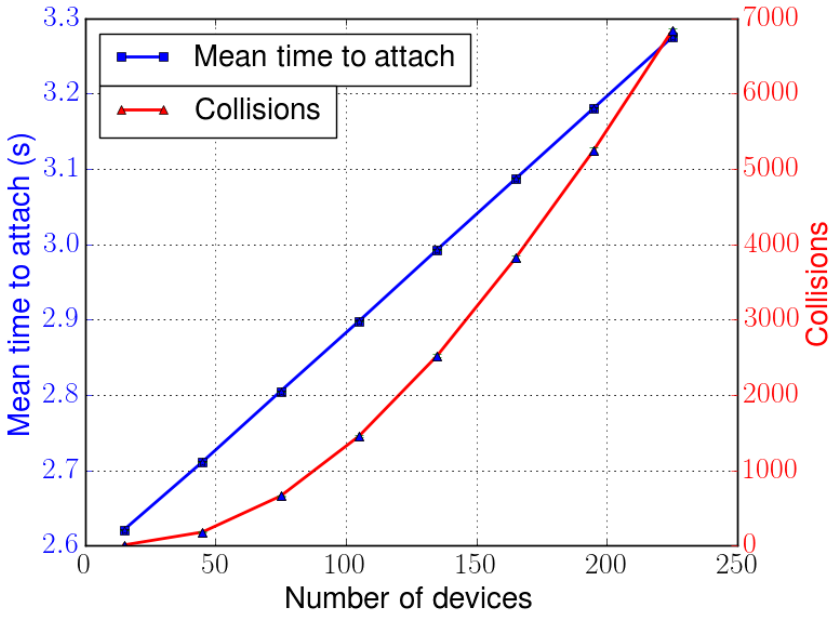

Figure 3.16: NB-IoT: Mean time to complete the RA procedure (left y-axis), total number of collisions in the first preamble message (right y-axis); with increasing number of UEs simultaneously sending access requests.

RAR Window Size The effect of RAR (i.e. Random Access Response) window size on the mean RA completion time while keeping other parameters fixed is also studied, simulated and presented here. RAR window size determines the time between preamble transmission and reception of RAR message. This window size can be the number between 0 and 10 in the unit of subframes. Figure 3.17 shows the impact of different values of RAR window size on the mean RA completion time. There is a considerable decrease in the mean RA completion time with decreasing this parameter. This behavior is explained by the fact that with a shorter RAR window size, the UEs consider the RA procedure as failed if RAR is not received from eNB within this window. The failed RA procedures do not have any impact on the results here as in the simulator the mean time to attach is calculated for only those RA attempts that are successful.

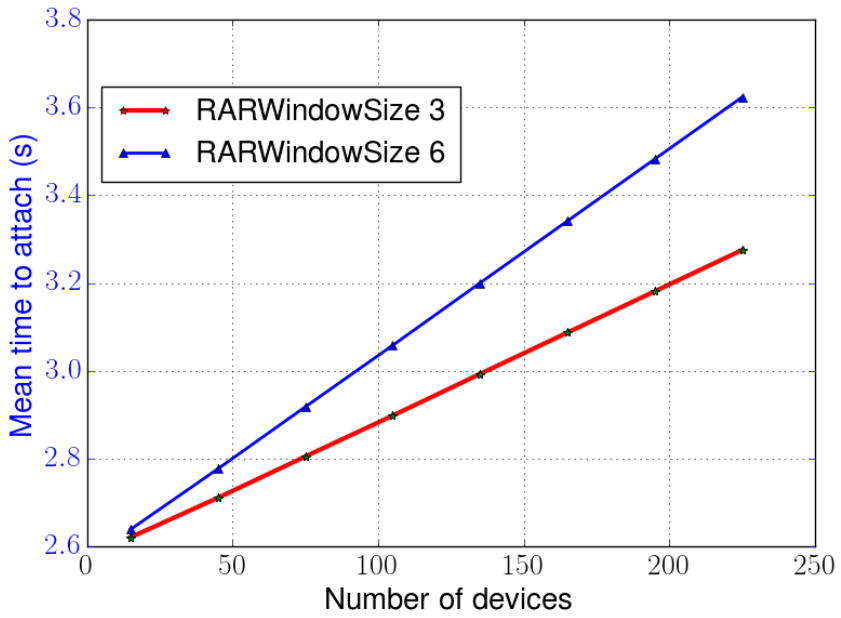

Figure 3.17: NB-IoT: Mean time to complete the RA procedure with different variants of RAR window size.

Maximum Preamble Transmissions The effect of *preambleTransMax* on the mean RA completion time while keeping other parameters fixed is simulated and presented in Figure 3.18. The maximum number of allowed transmissions for preamble are determined by this parameter. There is a considerable decrease in the mean RA completion time with more number of retransmission. The right choice of RA parameters can lead to improved performance from the system. However, RA procedure limits the minimum achievable latency of the system. The maximum preamble transmission of 25 and 50 exhibit similar results for mean time to attach. It is important to note that only those UEs retransmit that had a collision in the preamble transmission. As the preamble retransmissions increase, the probability of a collision decreases due to the fact that in each transmission cycle, there is a number of UEs that complete the RA procedure and only those facing a collision go for a retransmission. It has been noted that after 25 maximum preamble retransmissions, the majority of UEs have already completed their RA procedure and only a small number of UEs remain for retransmissions.

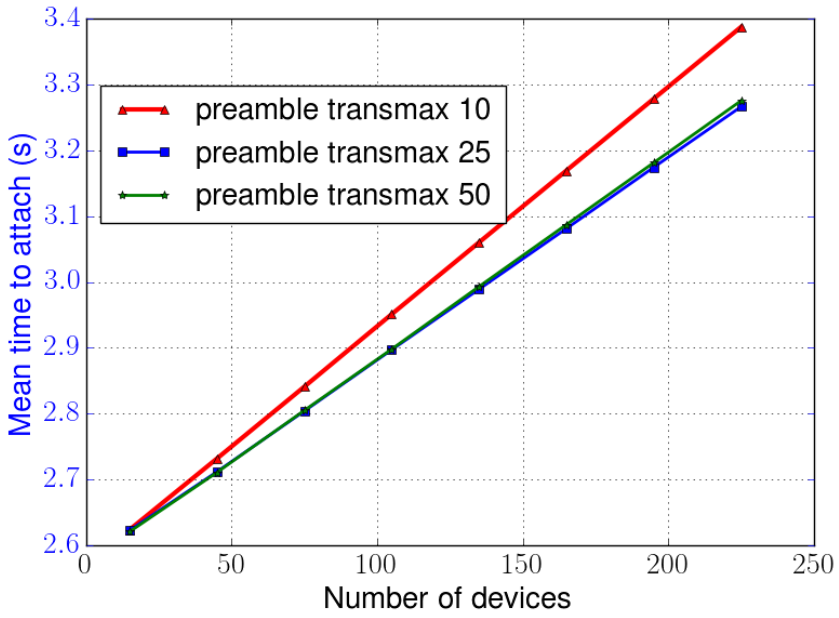

Figure 3.18: NB-IoT: Mean time to complete the RA procedure with different numbers of maximum preamble transmissions.

3.5 Conclusion

The uplink latency in LTE consists of two parts, control-plane and user-plane latencies. In the control plane, the contention-based random access operation is based on ALOHA-type access, which implies that, in case of simultaneous access requests from a massive number of devices, the network performance may degrade and some latency-critical application may be compromised. For example, in case of a vehicle collision on the road, a large number of vehicles would try to access the network in order to send/receive traffic information. In such scenario, the devices transmitting simultaneously may experience larger latency due to collisions on the preambles. Such application requires a seamless transfer of data in order of 10 ms while current LTE network may not be able to provide such low latency. On the other hand, user-plane latency consists of multiple components and the minimum achievable uplink user-plane latency in LTE is higher than 10 ms. This also means that LTE, in its current standards, cannot support URLLC applications.

The simulation evaluation of LTE control plane latency indicate that in case of simultaneous access requests, the latency can go up to several seconds depending on the control plane RA procedure configurations. The UEs that remain in connected state and want to send the data in uplink, can achieve a latency between 12 ms to 15 ms in case the network resources are not saturated. This latency could meet relaxed latency requirements of traffic efficiency or process automation use cases mentioned in Table 1.2. However, for other latency-critical use cases, LTE latency is higher and need to be reduced by means of different improvement techniques.

In this chapter, the LTE user and control plane latencies were theoretically explained and then compared through realistic simulations. The simulator design was also presented. The components that contribute to the uplink latency have been identified. This chapter evaluates the latency of LTE networks and opens up the question of latency reduction in LTE. The next chapter covers this question and presents some possible answers.

4

Latency Reduction in 4G Cellular Networks

4.1 Résumé

L'évolution vers la 5G est motivée par les nouveaux cas d'utilisation de l'IdO où, pour certaines des applications, une très faible latence est nécessaire, ce que les réseaux LTE actuels ne permettent pas. À cette fin, à partir de la version 13 du projet de partenariat de troisième génération (3rd Generation Partnership Project/3GPP) des techniques de réduction de la latence qui pourraient être intégrées aux réseaux cellulaires 5G ont commencé à être proposées et étudiées. Les travaux disponibles dans la littérature sur la réduction de la latence dans les réseaux 4G sont décrits en détail dans ce chapitre. L'étude bibliographique présentée se concentre en particulier sur deux techniques de réduction de la latence appelées Short Transmission Time Interval (sTTI) et Semi-Persistent Scheduling (SPS). Les domaines non explorés dans la littérature sont mis en évidence et ouvrent la voie pour l'utilisation des réseaux cellulaires pour des applications MTC actuelles et futures. Dans ce chapitre, ces techniques de réduction de la latence (TTI courts et SPS) sont analysées et leur potentiel à supporter des applications à faible latence est évalué. Ces techniques ne pouvant être évaluées dans aucun simulateurs de réseau open source, ces deux techniques ont été implémentées, puis évaluées dans le simulateur ns-3 (Network Simulator – 3). L'évaluation couvre les domaine applications d'automatisation industrielle avec des modèles de trafic périodiques et des modèles de trafic déclenchés par des événements, ainsi que des applications V2X avec des téléchargements périodiques de données. Les paramètres des temps de transmission (TTI) de 2, 7 et 14 (le

standard) symboles OFDM (2-os/7-os/14-os) ainsi qu'un ordonnancement dynamique et semi-persistant ont pris en évalués. Les différentes longueurs d'intervalle de transmission (TTI) et le processus d'ordonnancement semi persistant (SPS) implémentés dans le simulateur open source ns-3 ont été évalués pour les catégories d'équipements utilisateur (UE) LTE-M et IdO à bande étroite (Narrow Band IoT - NB-IoT). Les résultats montrent que, pour un seul équipement utilisateur (UE), un TTI court de 2-os avec un ordonnancement SPS réduit la latence de plus de 85% par rapport au TTI standard de 14-os avec un ordonnancement dynamique. Avec un nombre accru de UE, où certains UE envoient des données périodiquement et d'autres sporadiquement, les résultats montrent qu'un TTI court de 2-os avec un ordonnancement SPS ou DS peut réduire considérablement la latence. Ainsi, ces combinaisons ont le potentiel de prendre en charge les applications URLLC avec des exigences de latence strictes.

Le TTI court et le SPS peuvent réduire considérablement la latence au prix d'une augmentation des messages de contrôle et de l'utilisation des ressources. En outre, la prise en charge de TTI de tailles différentes au sein d'une même cellule est un problème complexe et doit être évaluée plus avant. Il est évident que les UE qui ne peuvent utiliser que le TTI standard ne peuvent pas utiliser les sTTI. Par conséquent, assurer la rétrocompatibilité des TTI est une étape nécessaire pour permettre la mise en œuvre de ces techniques dans des réseaux cellulaires 5G à très faible latence.

4.2 Introduction

The emerging MTC applications pose new challenges for the cellular communication systems to keep up with the very low latency and high reliability requirements [5]. Current LTE standards do not provide support for such use cases mentioned in Table 1.2. To meet the requirements of these demanding applications, there have been a few proposals for latency reduction techniques [77–79]. In this chapter[1], some of the techniques proposed for reducing LTE latency are outlined. A comprehensive analysis of literature survey for two of these techniques is also presented. These techniques were developed for ns-3 LTE module for simulations and evaluated over realistic URLLC use case scenarios. A detailed description of the development is also presented.

[1] Versions of this subsection have been published as original conference papers [23, 80–83].

- Author did the literature review, implemented the techniques in ns-3 LTE module, evaluated the implementation and prepared figures for the publications.
- Author also prepared the manuscripts for [80, 81].
- K. Nsiah and author prepared the manuscripts for [23, 82, 83].
- A. Sikora, B. Hilt, and JP. Lauffenburger gave fruitful advices and reviewed the articles.

4.3 Techniques for Latency Reduction - Introduction and Literature Survey

The control-plane and user-plane latencies analyzed in Chapter 3 highlight the requirements for reducing these latencies in cellular networks. To meet the requirements of low latency IoT applications, 3GPP proposed latency reduction techniques in Release 13 [77]. The improvements from these techniques target user-plane latency both in uplink and downlink. Among the techniques, short Transmission Time Interval (sTTI) and Semi-Persistent Scheduling (SPS) are selected and investigated in this thesis.

The existing wireless technologies have not been able to fulfill stringent requirements of QoS for industrial automation applications, therefore, these applications heavily rely on the wired fieldbus standards. Thus the use of wireless technologies in industry automation is limited to monitoring applications [65]. Currently used wired fieldbus standards for industry automation include: PROFINET, HART, SERCOS and CAN [88].

There have been continuous enhancements in LTE standards by 3GPP through different releases [8]. The LTE user-plane latency remained unchanged from 3GPP Release 8 till Release 13 and does not meet the requirements of URLLC use cases. As part of the evolution of the fourth generation LTE networks, different latency reduction techniques have been described and evaluated in 3GPP study and work items [77], [78], [79]. Some of these techniques have been mapped to the different parts of the cellular network architecture and shown in Fig. 4.1. For example, mobile edge computing, which enables offloading of RAN and EPC functionalities in powerful computing units in close proximity to the UEs, overlaps between the RAN and EPC. On the other hand, short Transmission Time Interval (sTTI) overlaps between the UE and eNB and enables shorter transmission times between both. Semi-Persistent Scheduling (SPS) [80] and Fast Uplink Access (FUA) [89] targeting latency reduction on the MAC layer were the first steps for latency reduction in Release 14 [78]. On the PHY layer, short TTI and reduced processing time belong to the second step towards latency reduction included in Release 15. In the following, an introduction and the research works analyzing the potential of latency reduction techniques are outlined.

4.3.1 Short Transmission Time Interval

As presented in section 3.3.1, the Transmission Time Interval (TTI) duration in LTE is 1 ms with two slots of 0.5 ms each, comprising 14 OFDM symbols (14-os) of 0.0714 ms each. With a reduction of the TTI length, the overall data transmission and processing time can be reduced significantly [90,91]. As presented in [92], short TTI of 2 OFDM symbols (2-os) lasting 0.14 ms

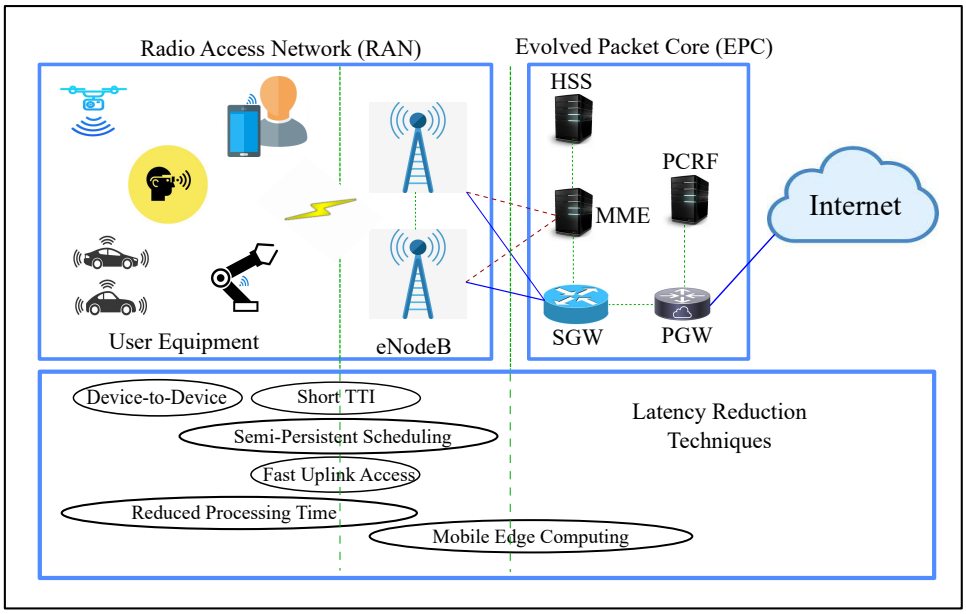

Figure 4.1: Mapping of different latency reduction techniques to cellular network architecture.

and one slot of 7 OFDM symbols (7-os) lasting 0.5 ms shall be standardized in Release 15 to be supported by downlink and uplink transmissions. The reduced TTI implies shortened time duration for transmission (see Figure 4.2) and Short Processing Time (SPT) as well. TTI shortening can also lead to a backward compatibility issue as all the control channels are designed for 1 ms TTI. For instance, Downlink (DL) control channel in legacy LTE occupies 1-3 symbols, which would require modifications to enable sTTI of two symbols. UEs supporting sTTI shall be able to coexist with legacy UEs within the same system bandwidth. This can be achieved by a flexible frame structure allowing both types of UEs to transmit. Furthermore, sTTI can also lead to a decreased throughput as with sTTI the amount of resources required for control channels also increases [92]. In the following, the investigation available in literature for latency reduction with sTTI are outlined.

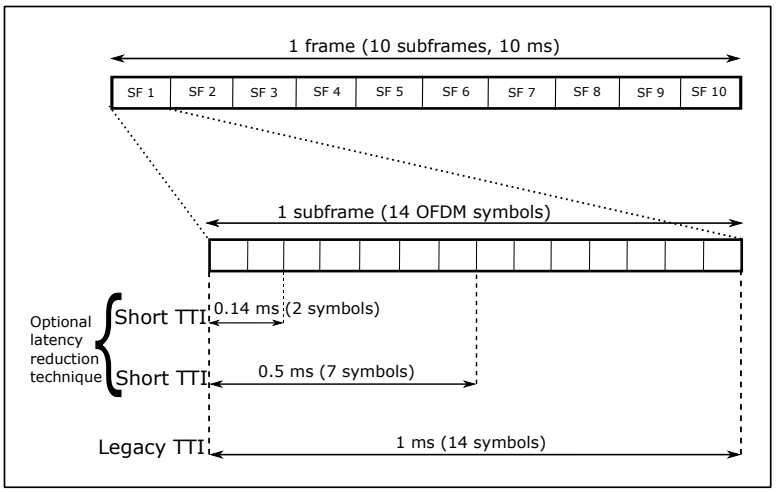

Figure 4.2: LTE frame structure: A subframe consists of 14 OFDM symbols/resource elements [93]. Transmission time interval in legacy LTE is based on one subframe i.e. 1 ms. The short TTI could be used as 7-os or 2-os.

In [94], Hosseini et al. conduct a link-level performance analysis of low latency operations in the downlink and uplink of LTE networks with different shortened TTI lengths of 1-os, 2-os, 7-os and legacy 14-os. The authors perform simulations on a 10 MHz wideband UE category with an in-house developed simulator. Their work mainly discusses the backward compatibility for UEs that do not support sTTI. There are two shortcomings in their work, i) the evaluation of short TTI is conducted for wideband UEs and narrowband UE category is not addressed, ii) results are lacking in confidence because there is no information provided about the used simulator and its degree of realism.

Arenas et al. [90] conduct a performance analysis of the potential benefits of the short TTI and reduced processing time techniques for the reduction of the downlink and uplink latencies as well as the round trip time latency in LTE networks. Short TTI of 2-os and 7-os are used in their simulations and are evaluated for eMBB use cases. Their results show that sTTI outperforms reduced processing time with respect to uplink, downlink and round trip latencies. The focus of their work is the reduction of latency for eMBB use cases. Therefore, latency reduction for URLLC use cases is not addressed. A Performance evaluation of the combined short TTI and reduced processing time techniques are also missing. As for [94], techniques for narrowband category of UEs are not evaluated.

Xiaotong et al. [95] proposes a combination of 7-os TTI together with the reduction of uplink access delay. The main idea behind their work is to reduce the processing time both at eNB and UE according to the TTI length. The performance analysis of the system latency of the different TTI lengths and their proposed technique is evaluated for cell center and cell edge users by running different simulation scenarios. Thus, their work focuses on improving latency for cell center and cell end-users. However, the evaluation does not address the narrowband UE category. Furthermore, the paper provides no explicit information concerning the type of simulator used.

Two latency reduction techniques, reducing processing time and shortening the TTI were studied separately in [96]. In order to achieve a comparatively smaller latency, short TTIs of length 1-os, 2-os, 7-os, 14-os were used. The maximum time advance and transport block size are used to shorten the processing time. The authors claim that short processing time with an unchanged frame structure results into less standardization efforts. However, it is hard to support strict requirement on latency such as less than 1 ms with only short processing time. Short TTI length is taken into account and 2-symbol length can achieve a trade-off between overhead and performance gain. Comparing the TTI lengths, 1-os offers the best performance in terms of latency reduction; however, this introduces a significant overhead. An Evaluation

of the technique for narrowband UEs is not considered. Finally, the major contribution focuses on the analysis of the proposed design and lacks the validation by simulations.

Performance improvement of short TTI is also evaluated in the downlink of LTE networks in [97]. The proposed latency reduction model is based on the traffic arriving rate and the TTI length. The authors also propose a downlink scheduling algorithm to take the impact of shortened TTI into consideration. In their simulations, the TTI of 2-os, 4-os, 7-os and 14-os were used. Their results show that the shortest TTI configuration (i.e. 2-os) offers the best performance with regard to the latency reduction at low traffic loads. Two areas are considered, the evaluation of short TTI in the uplink of LTE and the evaluation of the technique for narrowband category of UEs. The evaluation of their proposed techniques are based on an in-house developed system-level simulator. They use a wideband UE category with 10 MHz bandwidth. Therefore, the impact of latency reduction techniques on narrowband UE category is not investigated.

Aktas et al. [65] evaluate short TTI for industrial automation applications to identify the set of use cases that could be supported with the latency reduction enhancements. They evaluated sTTI also in a system-level in-house simulator over industry automation parameters. The results show that a combination of sTTI with reduced processing time can reduce the end-to-end latency to 3 ms. However, their work lacks the evaluation of latency reduction techniques for narrowband UE category.

The latency of cellular-based V2X with shortened TTI was analyzed and verified by Lee et al. in [98]. To verify the feasibility of V2X services, the V2X latency was divided into two types of latency, TTI-independent latency and TTI-proportional latency. The latency of cellular-based V2X systems was evaluated using system-level simulations considering additional overhead from shortened TTI. They evaluate 14-os, 7-os and 1-os TTI for V2X-based applications. The results show that 1-os TTI can support V2X application that require less than 10 ms latency. However, the evaluation is based on 6 GHz frequency band with a 10 MHz bandwidth UE. The evaluation of narrowband UE is not covered in their work. Moreover, the simulations do not cover sTTI evaluation on the legacy LTE bands.

Fehrenbach et al. discuss the challenges and design problems for low latency cellular networks in [99]. The latency reduction techniques including sTTI to be standardized in 3GPP Release 15 are analyzed. Theoretical model and calculations mainly covering the physical layer aspects are proposed. Validation of the techniques through realistic simulations for URLLC specific use cases is not provided. Li et al. [92] conduct a link-level performance analysis of low latency operation in the downlink and uplink of LTE networks with different TTI lengths of 1-os, 2os, 7-os and compare the results with legacy 14-os. The paper analyzes the challenges

of channel estimation and data demodulation associated with short TTI lengths. This is because short TTIs may not include the Common Reference Signals (CRS) and Demodulation Reference Signals (DMRS), which leads to inferior channel estimation performance compared to legacy TTIs. As a result of these challenges, their work focuses on investigating several link level analyses, which includes using different CRS and DMRS configurations, different placements of reference signals within each shortened TTI etc. to ensure a trade-off between channel estimation quality and transmission gains. The evaluation of short TTI for narrowband category of UEs is not covered. Their results show that the low-latency LTE network with different TTI lengths outperform legacy LTE in both downlink and uplink under different operating scenarios (e.g. transmission modes, channel models etc.).

The performance of cellular networks has been investigated through system-level simulations in a realistic factory deployment scenario by Ashraf et al. in [28]. In their simulation configurations, TTI of 1 ms and 0.2 ms were used along with short processing time on both UE and eNB sides. Their results show that the reduced TTI in a factory hall configuration of network, outperforms legacy TTI in terms of uplink latency. They also discuss and compare analytical and simulated latency numbers. However, the bandwidth used in the simulations is 5 MHz, and therefore does not represent the improvements and impact of sTTI on a narrow bandwidth system.

Pocovi et al. investigated PHY and MAC layer enhancements for LTE networks to support URLLC use cases. They analyzed 14-os, 7-os, and 2-os TTI and concluded that there is a significant benefit of using short TTI size as it reduces the over-the-air transmission delay, frame alignment, and HARQ retransmission time. The sTTI improvements were verified by an in-house developed simulator with 10 MHz system bandwidth and other LTE parameters. The control overhead associated with sTTI was also investigated and it was made clear from the results that sTTI has a larger overhead as compared to legacy 14-os TTI. However, as for the most of the references cited above, the performance evaluation of low-cost narrowband UE is not covered. In [100], Pocovi et al. discuss the challenges faced to enable URLLC applications and propose different solutions covering multiple aspects of the radio interface. System-level simulation results are presented, showing how the proposed short TTI technique can work in order to fulfill the ambitious latency and reliability requirements of upcoming URLLC applications. The simulation results highlight the importance of using short TTI in combination with fast HARQ retransmission mechanisms for efficient scheduling of the latency-critical payloads. In this regard, early prediction of the HARQ feedback is suggested in order to reduce the processing time during retransmissions and fulfill the 1 ms URLLC

latency budget. Their work lacks the evaluation of low-cost narrowband UEs and the realistic simulations with URLLC use case parameters.

A summary of above analyzed proposals from the literature is presented in section 4.3.4.

4.3.2 Semi-Persistent Scheduling

In LTE, coordination of the radio resources and channel access is done by the eNB. However, conventional dynamic scheduling includes extra control overhead for uplink transmission. As depicted in Fig. 4.3 (left), a UE needs to get access to uplink resources by sending a scheduling request first. The eNB then allocates resources for the UE through scheduling grant. As presented in [9], with a default TTI of 1 ms, the uplink latency in LTE networks is always above 12 ms. SPS overcomes the extra control messages delay by eliminating the scheduling messages. As shown in Fig. 4.3 (right), a UE does not have to initiate a scheduling request and can transmit the data as soon as it is ready to be sent out. With a default TTI of 1 ms, SPS can reduce the uplink latency to 4.5 ms with a cost of reduced capacity due to the resource pre-allocation [80].

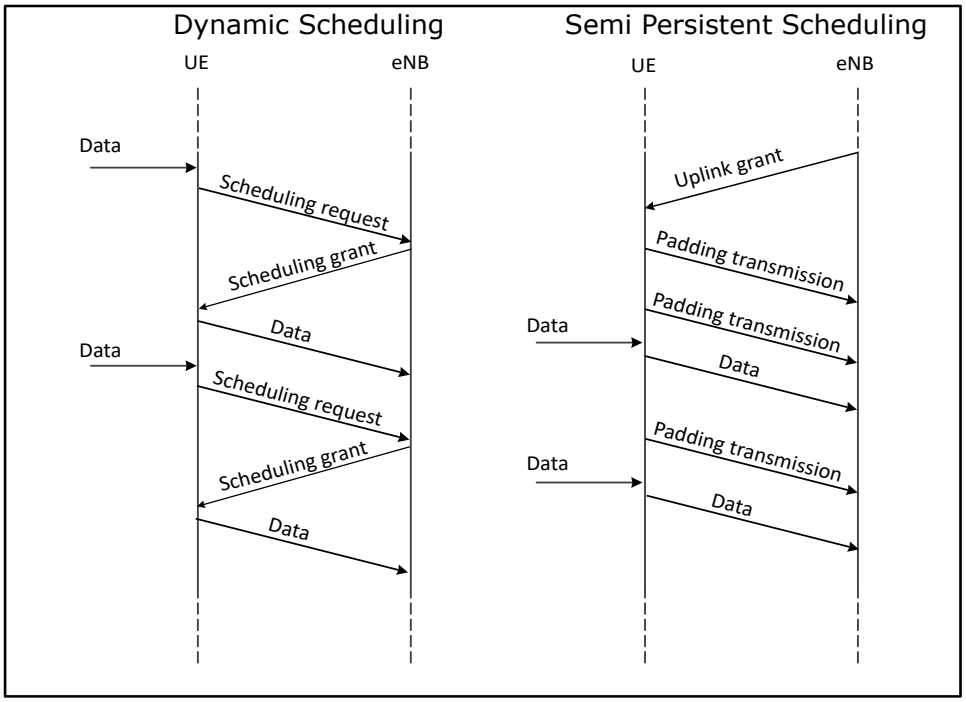

Figure 4.3: Resource scheduling types in LTE, (left) Dynamic scheduling: UE sends a scheduling request when data is ready for uplink and in turn receives a scheduling grant for uploading data, (right) Semi-persistent scheduling: The eNB allocates resource to the UEs on an a-priori basis for periodic data uploads. The UEs send padding information when there is no data to send.

LTE systems usually use Dynamic Scheduling (DS) to maximize resource utilization where the UEs send scheduling requests to get access to the network resources for uplink data transmissions. In SPS, the eNB schedules uplink transmissions for the UEs in connected state without the reception of scheduling request. As a UE does not necessarily have data

available to send, SPS becomes inefficient due to spectral utilization. In such case the UE sends padding information in the subframes/symbols scheduled for it (see Fig. 4.3) and wastes the network resources. This method incurs two issues into the system, a) spectral inefficiency and b) increased energy consumption. Both issues are due to the unnecessary transmissions in the scheduled resources when the UE does not have any data to transmit.

To understand the improvements these latency reduction techniques offer, the analysis of latency components presented in Section 3.3 is further extended with sTTI and SPS in Table 4.1. Both short TTI variants (7-os and 2-os) reduce the latency of three components where data transmission is made accordingly. On the hand, SPS eliminates the scheduling request/grant procedure and reduces the latency from respective components. A combination of both sTTI (with 2-os) and SPS reduces the uplink latency by more than 85% as per the theoretical calculations.

Table 4.1: Theoretical calculations for uplink latency in milliseconds with latency reduction techniques. 14-os with dynamic scheduling is the baseline legacy LTE scheme where all other variants are from the latency reduction techniques.

Component	14-os DS	14-os SPS	7-os DS	7-os SPS	2-os DS	2-os SPS
Average delay to next scheduling request opportunity	0.5	0	0.5	0	0.5	0
SR transmission	1	0	0.5	0	0.14	0
eNB processing delay	3	0	1	0	0.43	0
Transmission of scheduling grant	1	0	0.5	0	0.14	0
UE processing delay	3	3	1	1	0.43	0.43
Average delay to scheduled resources	0.5	0.5	0.5	0.5	0.5	0.5
Transmission of uplink data	1	1	0.5	0.5	0.14	0.14
eNB processing delay	1.5	1.5	1	1	0.43	0.43
Total uplink latency	**11.5**	**6**	**5.5**	**3**	**2.71**	**1.5**

*DS: Dynamic Scheduling, SPS: Semi-Persistent Scheduling, os: OFDM Symbol

There have been only a few works in literature investigating the improvements in latency using SPS in LTE networks. In [65], Aktas et al. evaluate SPS for latency sensitive industrial automation applications. The simulations were performed over an area of 80 m^2 with 10 MHz system bandwidth. The results from the periodic traffic simulated in an in-house developed simulator show that SPS can support low latency use cases.

Arenas et al. [90] also evaluated SPS to show that improved LTE can support novel low-latency use cases. Their simulations performed on a wideband UE category show that

using SPS with sTTI can reduce round trip time from 16 ms to 3 ms in ideal conditions (i.e. without the consideration of channel characteristics). Similar to other works, they also do not investigate narrowband UE category.

One of the potential problems with SPS is that it increases the resource utilization inaccuracy. In certain cases, when UEs do not have any data to transmit, the allocated resources are wasted, thus decreasing the resource utilization efficiency. Therefore, it is also important to investigate the performance of SPS for narrowband UE categories, such as LTE-M and NB-IoT.

The performance of a proposed predictive allocation scheme based on SPS technique is evaluated in [101] for typical delay-sensitive process monitoring application scenarios. The proposed scheme additionally addresses the possible low spectrum utilization associated with SPS due to unused reserved resources. This is achieved by reserving resources for a subset of end-devices selected by the base station based on supervised learning of the correlation between the devices. The contribution of this paper is in two-fold: present SPS as a potential technique to enable mission-critical applications in industrial systems and address the issue of low spectrum utilization. Results show that pre-allocating radio resources to end devices reduces access latency compared with dynamic scheduling. The evaluation of their work for narrowband category of UEs is missing.

In [102], Afrin et al. evaluate the performance of dynamic and SPS scheduling types for LTE networks using the OPNET network simulator. Their work shows that SPS outperforms dynamic scheduling as it overcomes the excessive control channel overheads associated with dynamic scheduling. In dynamic scheduling, the UEs send a scheduling request before sending the actual data. They receive a response from eNB that contains the information about scheduled uplink resources where the data can be sent. Although SPS performs well for applications with periodic fixed-sized packet flows, it does not efficiently support applications with sporadic traffic patterns. The authors therefore propose an adaptive semi-persistent scheme that adjusts the allocation of resources based on traffic patterns and buffer filling reports. Results show that the proposed scheme performs best compared with dynamic scheduling in terms of satisfying delay requirements of M2M traffic. Their work focuses on wideband LTE networks and do not cover narrowband networks. Moreover, they do not address the use cases of industrial automation communication. Ashraf et al. [28] analyzed the performance of latency reduction techniques including SPS for URLLC use-cases. The MAC layer enhancements shown in their factory automation simulations results seem very promising. However, the evaluation of low cost narrowband UEs is not covered and it remains unclear how these techniques behave with shorter system bandwidth.

Ali et al. [103] investigated the challenges and opportunities for adopting the fast uplink grant to support MTC. To overcome the potential challenges, a two-stage approach that includes traffic prediction and optimized scheduling is proposed in their work. For this approach, various solutions for source traffic prediction for periodic machine type devices traffic are reviewed and methods for event-driven traffic prediction are proposed. For optimal allocation of uplink grants, they also proposed new solution based on advanced machine learning methods. The interesting fast uplink access technique proposed in their work lacks the validation through realistic simulations. It is also unclear that how the proposed technique works for narrowband UE categories.

Abreu et al. [104] proposed a scheme, in which a group of users shares a pre-scheduled resource for retransmission to overcome the grant scheduling retransmission issue of conventional SPS. The benefit is that it provides a retransmission opportunity without needing a scheduling control information. Besides, if the pre-scheduled resource cannot be re-allocated, the sharing mechanism avoids excessive capacity loss. It is demonstrated through a simple analytical model that, for right group size and initial transmission error rates, an excellent target error probability can be achieved. It is also shown that the suggested scheme can provide improved resource efficiency compared to a single conservative transmission, which also avoids re-scheduling. However, their work does not evaluate the proposed scheme in a realistic simulation environment. Furthermore, different system bandwidths are not investigated as well.

4.3.3 Other Latency Reduction Techniques

There are some other latency reduction techniques that were initially proposed by 3GPP [77] and later investigated in different works. These techniques include reduced processing time, fast uplink access and hand-over without random access. Short processing time has been investigated by all the works in literature that evaluate sTTI. In the previous subsection, all papers that have worked towards sTTI evaluation, also use short processing time. On the other hand, FUA is an extension of SPS where UEs do not send any padding information in the uplink when there is no data available to send. There is no difference in SPS and FUA in terms of latency as both eliminate the dynamic scheduling overhead. However, FUA is more energy efficient since there are no padding transmissions in the allocated resources where UE does not have any data to send in uplink.

The author has considered sTTI and SPS as the latency reduction techniques for this thesis due to the fact that both are applicable for industry automation and V2X use cases. Other techniques either work for one of the use cases or they are similar to SPS in terms

of latency. For instance, as stated earlier, FUA has a similar mechanism as SPS for latency reduction. Mobile edge computing is a concept mainly used for campus networks and therefore not considered for V2X use cases. It does not seem practical to install local computing nodes on each base station in a cellular network. Random-access less hand-over also reduces latency according to [77]. The hand-over procedure where a UE moves from the serving area of one eNB to another, is not supported in NB-IoT. Moreover, for industry automation use case, the nodes are normally not moving (see Table 1.2). Therefore, random access less hand-over technique is not effective for industry automation use case.

4.3.4 Summary

Summarizing the contributions for latency reduction, short TTI, SPS, and reduced processing time techniques are classified and presented in Table 4.2. It can be noted from the presented table that there are three topics that have not been studied in the literature.

- The evaluations of latency reduction techniques were mainly conducted for broadband UE categories. Evaluation of narrowband UE categories (i.e. LTE-M and NB-IoT) are missing.
- Evaluations were conducted using either in-house simulators or theoretical calculations. The latency reduction techniques are not implemented and evaluated in any open-source simulators. There is no open-source implementation available for any of the techniques, which makes it difficult for the research community to evaluate and enhance the techniques.
- Simulation evaluation of latency reduction for realistic use case scenarios is also not present in the literature. In 5G ecosystem, different use cases pose different type of challenges, therefore, it is also important to investigate the techniques with realistic use case scenarios. For example, an evaluation of V2X use cases requires most of the nodes in the network to be mobile such as vehicles, whereas for industry automation that is not required.

The work in this thesis is directed towards filling the gap presented above. A complete analysis and optimization of sTTI and SPS for narrowband UE categories is presented and evaluated with a widely used open source simulator.

Table 4.2: Classification of contributions by different papers in the literature for latency reduction in cellular networks.

Paper ref.	BW	Target use case	Eval. platform	NB eval.	Use case eval.	sTTI	SPS	SPT
[28]	5 MHz	URLLC	in-house	✖	✔	✔	✔	✔
[65]	10 MHz	URLLC	in-house	✖	✔	✔	✔	✔
[90]	10 MHz	eMBB	in-house	✖	✖	✔	✖	✔
[92]	10 MHz	-	in-house	✖	✖	✔	✖	✖
[94]	10 MHz	-	in-house	✖	✖	✔	✖	✖
[95]	20 MHz	-	in-house	✖	✖	✔	✖	✔
[96]	20 MHz	URLLC	theory	✖	✖	✔	✖	✔
[97]	10 MHz	URLLC	in-house	✖	✖	✔	✖	✔
[98]	10 MHz	V2X	in-house	✖	✔	✔	✖	✖
[99]	-	-	theory	✖	✖	✔	✖	✖
[100]	10 MHz	URLLC	in-house	✖	✖	✔	✖	✔
[101]	-	URLLC	theory	✖	✖	✖	✔	✖
[102]	3 MHz	URLLC	OPNET	✖	✖	✖	✔	✖
[103]	-	MTC	theory	✖	✖	✖	✔	✖
[104]	-	URLLC	theory	✖	✖	✖	✔	✖
[105]	10 MHz	URLLC	in-house	✖	✖	✔	✖	✖

*BW: Bandwidth, *NB: Narrowband, *SPT: Short processing time

4.4 Development of Latency Reduction Features in ns-3

The current implementation of the LTE module in ns-3 does not support any of the latency reduction techniques. In order to evaluate them, SPS and sTTI feature for LTE module are implemented in this thesis to include TTI of 0.5 ms (7-os) and 0.14 ms (2-os) along with the legacy TTI of 1 ms (14-os). The processing time reduction is also implemented with the shortened TTI. The following subsections describe the implemented features.

4.4.1 Short Transmission Time Interval: ns-3 Implementation

In the implementation of ns-3 LTE module, the transmission time interval and length of subframe were fixed to 1 ms as the concept of sTTI was not initiated yet. Therefore, both the TTI and subframe length were controlled by only one variable with a value of 1 ms. The

first challenge in the development of sTTI was to isolate the implementation of both of these parameters. It is also important that a user can configure the default network TTI to any of three options (2-os, 7-os, 14-os) for the simulations. Since TTI is related to the PHY layer of LTE protocol stack, most of the development is based on the PHY and Spectrum modules of ns-3 LTE. The transmission of data packets is performed in the *lte-ue-phy* and *lte-enb-phy* classes for UE and eNB respectively. In the simulator, the transmission is made when 14 OFDM symbols are encoded and available at the layer. This model has been changed in our implementation and the transmission is now possible according to the set value of TTI, which could be 14-os, 7-os, or 2-os. A generic overview of the simulation model is given in Figure 4.4.

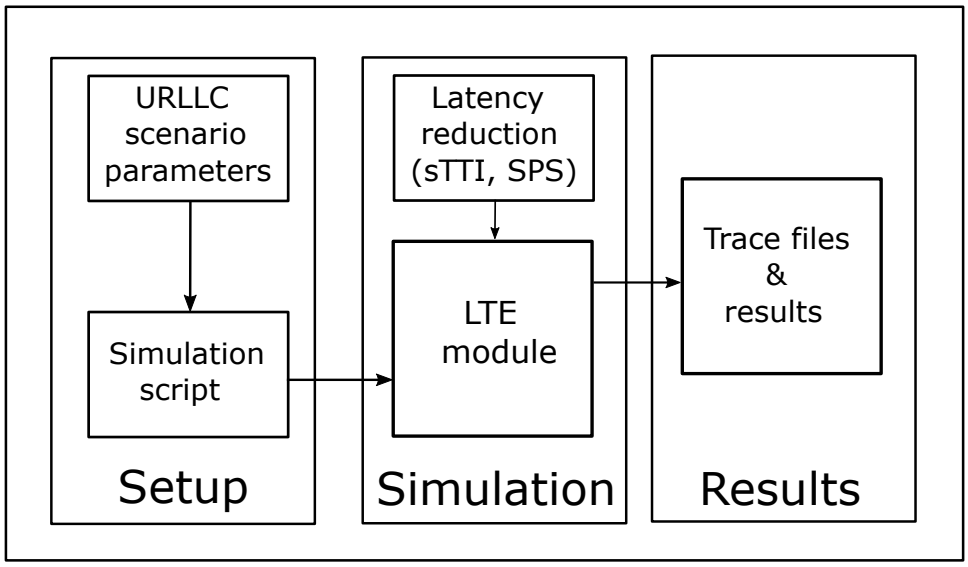

Figure 4.4: A block diagram of simulation setup in ns-3. The latency reduction techniques are developed and LTE model with realistic URLLC scenario parameters was evaluated.

4.4.2 Semi-Persistent Scheduling: ns-3 Implementation

The MAC schedulers available in ns-3 LTE module [106] are all based on the principle of dynamic scheduling. The SPS functionalities are added into the dynamic scheduler. The network resources are allocated to the connected UEs with the reception of SR by the eNB. The eNB considers those UEs connected that have completed the RA procedure. In the general MAC scheduler implementation, the eNB always receives the Buffer Status Report (BSR) from the UEs at the start of each subframe. However, the BSR becomes irrelevant in case of SPS, as the resources are allocated anyway to all the connected UEs. The implementation of SPS can easily be adopted for any other MAC scheduler available in ns-3 LTE module. The major changes, which were made to make the SPS scheduler work, are in the *StartSubFrameIndication* function of *RRMacScheduler* class. The scheduler checks if the UE is connected, and there are

enough resources available to schedule this UE. When all these conditions are met, the UE is informed about its pre-allocated resources.

In the following section, the simulation evaluation of both developed features (i.e. sTTI and SPS) is presented and results are discussed.

4.5 Experimental Evaluation of Latency Reduction Techniques

The latency reduction features proposed and implemented in this thesis work are evaluated for use case requirements mentioned in Table 1.2 (i.e. industry automation and V2X). The following subsections describe overall simulation setup and evaluation of latency reduction techniques for URLLC use cases. The primary aim of this evaluation is to observe the latency when different TTI are used and the number of UEs is growing both with dynamic and semi-persistent scheduling.

4.5.1 Simulation Setup

The parameters used in the simulations for this section are shown in Table 4.3. Simulations were performed for 3GPP LTE-M and NB-IoT UE categories with a bandwidth of 1.4 MHz (6 PRBs) and 200 kHz (1 PRB) respectively. In order to evaluate reduction in latency through sTTI and SPS, three scenarios of URLLC were evaluated to investigate the reduction in latency.

The first part of simulations presents a single stationary UE in the network and compares simulated latency numbers with theoretical calculations. In the second part, parameters for URLLC use cases are used and the latency reduction techniques are evaluated.

- For industry automation use case, hybrid building propagation loss model from ns-3 is used for realistic channel propagation model and UEs are placed inside a large hall in a building. The simulations are targeted towards evaluating process automation where UEs send data periodically.
- For V2X use case, the mobility of nodes on a small sized city map with realistic propagation loss model is evaluated. The work in this thesis only focus on network assisted V2X communication. Thus, the Device-to-Device (D2D) communication is not considered here.

In the third and last part of simulations, industry automation use case is evaluated to analyze and discuss the trade-offs between device density and uplink latency.

Table 4.3: Simulation parameters used in ns-3 simulations for the evaluation of uplink latency with short transmission time interval and semi-persistent scheduling.

Parameter	Value
Simulator - version	ns -3.26
Propagation loss model	Three log-distance / Hybrid building
eNB transmission power	43 dBm
UE transmission power	20 dBm
Uplink bandwidth	1.4 MHz (LTE-M), 200 kHz (NB-IoT)
Number of resource blocks	6, 1
Packet size	50 B
Packet Tx interval for periodic traffic	100 ms
Transmission time interval	{1, 0.5, 0.14} ms
Resource scheduling	Dynamic / Semi-persistent
Simulation duration	30 s

4.5.2 Results

4.5.2.1 Simulator Validation

As mentioned earlier, only a single UE sending and receiving data from the eNB is considered in the first part of simulations as shown in Fig. 4.5 to analyze the impact of latency reduction techniques on the uplink latency.

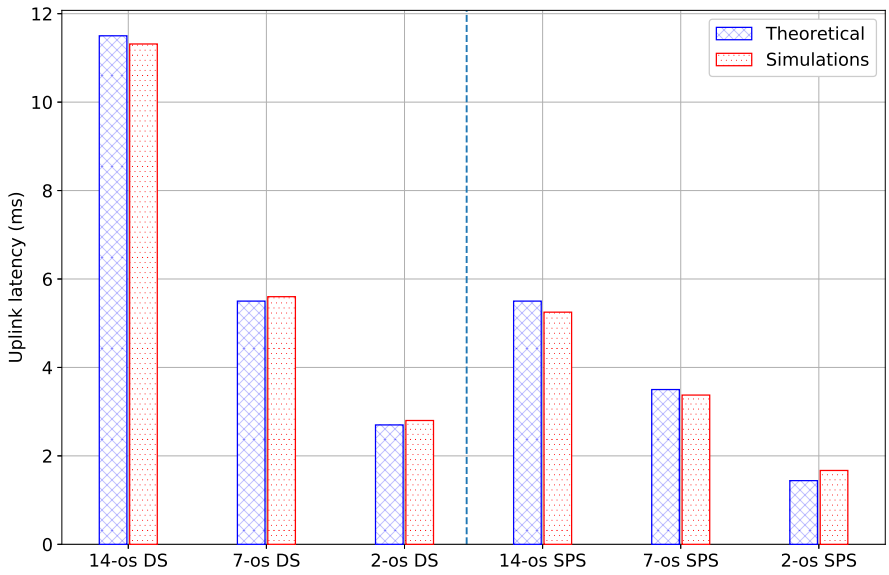

Figure 4.5: Uplink latency comparison of theoretical and simulated values for three different TTI and two different scheduling schemes with a single LTE-M UE in a cell sending data to eNB. TTI of 2-os with SPS reduces the latency for approximately 85% as compared to the baseline legacy 14-os TTI with dynamic scheduling.

The theoretical calculations of uplink, downlink and end-to-end latency with sTTI and SPS are presented in Table 4.4.

Table 4.4: Theoretical latency expectations for different lengths of short TTI together with dynamic or semi-persistent scheduling. The end-to-end latency includes only the access network delay. These numbers remain same for both type of narrowband UE categories i.e. LTE-M and NB-IoT

Feature	TTI (ms)	Scheduling	Processing time (ms)	UL latency (ms)	DL latency (ms)	E2E latency (ms)
14-os	1	Dynamic	3	11.5	4.5	16
14-os	1	SPS	3	5.5	4.5	10
sTTI 7-os	0.5	Dynamic	2	5.5	2.5	8
sTTI 7-os	0.5	SPS	2	3.5	2.5	6
sTTI 2-os	0.14	Dynamic	1	2.7	0.93	3.63
sTTI 2-os	0.14	SPS	1	1.44	0.93	2.37

A comparison between these theoretical and simulated uplink latency for a single UE is shown in Figure 4.5 for 2-os, 7-os and 14-os TTI with dynamic and semi-persistent scheduling. The small difference that we observe between simulated and theoretical latencies is due to the waiting time for the resources to send either scheduling request or data. As presented in Table 3.1, the average time to resources considered for theoretical calculations is 0.5 ms; however, due to randomization, it could be between 0 to 1 ms in simulations. Increasing the number of simulation runs would lead towards the average value of 0.5 ms. It is worth noticing that with 14-os TTI and dynamic scheduling, the minimum uplink latency is limited to 11.5 ms. Multiple transmissions required to complete a scheduling request/grant procedure (see Table 3.1) is the main cause for the minimum latency limit for 14-os TTI. In this case, each uplink and downlink transmission adds a delay of 1 ms. However, for shorter TTI, both the transmission and the processing times decrease, which results in a reduced overall latency. Moreover, SPS removes the scheduling request/grant messages, which further reduces the latency. The short TTI of 2-os together with SPS can reduce the uplink latency for a single device for more than 85% from the baseline legacy TTI with dynamic scheduling. The uplink latency for NB-IoT is also the same as for LTE-M (see Figure 4.6) since the user-plane messages follow the same structure in both cases. The comparison of theoretical and simulated uplink latency for both LTE-M and NB-IoT show that the development of latency reduction techniques is validated.

4.5.2.2 SPS and sTTI Evaluation for Industry Automation

The evaluation analysis of latency reduction techniques for LTE-M is further extended for ultra-low latency use cases in this subsection. Figure 4.7 shows the uplink latency for a varying

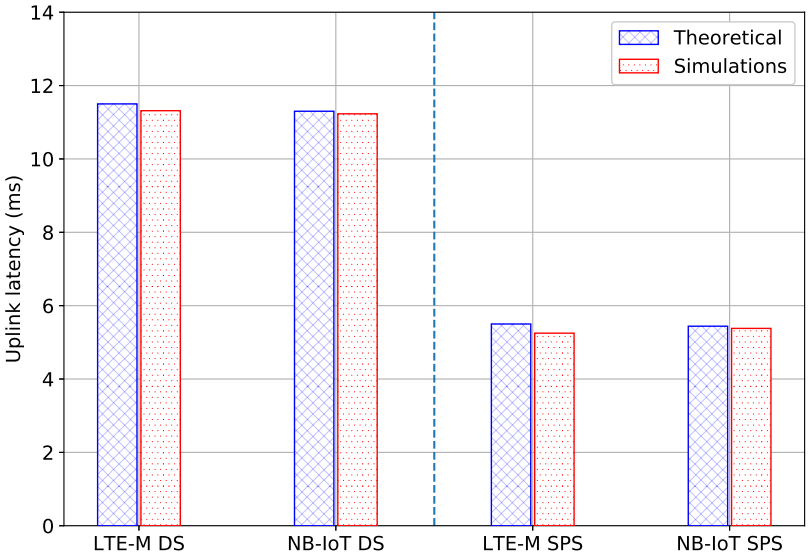

Figure 4.6: Uplink latency comparison between NB-IoT and LTE-M of theoretical and simulated values for two different scheduling schemes (i.e. dynamic and semi-persistent) with a single UE in a cell sending data to eNB.

number of UEs in a factory hall depicting the scenario of industry automation. All the UEs are considered to be stationary in the simulations, located randomly in the coverage area of the eNB, and periodically (i.e. 100 ms) send data to eNB. All UEs communicate with a single eNB installed in the factory hall. The periodic data transmissions represent mainly the process automation where sensors on the machines send data periodically. It is important to note from the results that short TTI can satisfy the 10 ms latency requirement for narrowband UE category with a limited number of UEs (i.e. <30) in a cell.

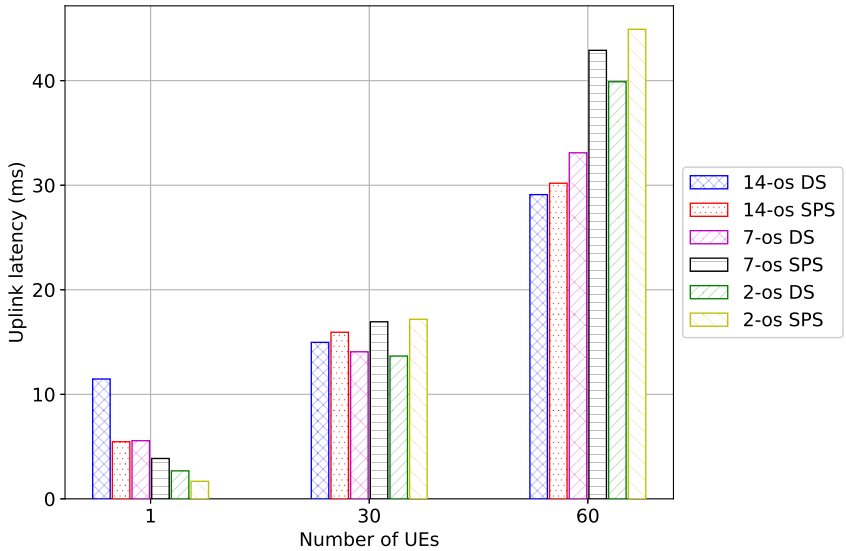

Figure 4.7: Uplink latency comparison of different TTI for industry automation use case where multiple static LTE-M UEs send data periodically to a single eNB.

The analysis of sTTI and SPS is further extended to discuss the trade-offs for satisfying URLLC latency requirements. Factory automation (event triggered data uploads) and process automation (periodic data uploads) are evaluated with different number of LTE-M UEs in the network cell. The rationale behind this evaluation is to discuss the limit on number of UEs while fulfilling the URLLC latency requirements. Fig. 4.8 represents the uplink latency for different number of UEs with periodic data uploads. The periodic data transmissions represent mainly the process automation where sensors on the machines send data periodically. The latency of legacy 14-os TTI with dynamic scheduling is always more than 10 ms, while shorter TTI (i.e. 2-os) with dynamic scheduling manages to keep the latency under 5 ms. The latency increase for SPS with higher number of UEs is due to the fact that the network resources are allocated for transmissions to all the UEs even if there is no data to send with the UEs. Therefore, SPS is less effective in case of higher number of UEs in a cell.

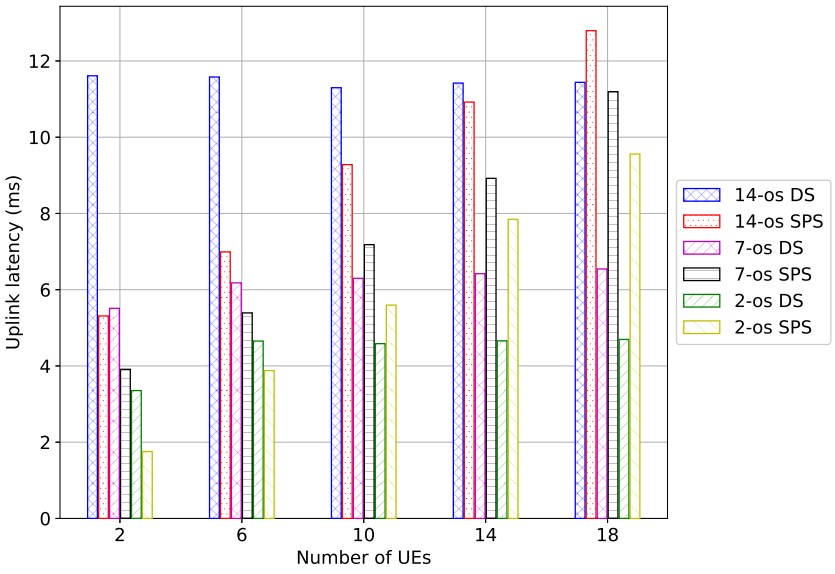

Figure 4.8: Uplink latency comparison of different TTI for industry automation use case with different number of LTE-M UEs sending date periodically to eNB. 2-os TTI is clearly a better choice over legacy 14-os TTI. The short TTI can further reduce the latency to 2 ms with help of SPS but only for a very small number of devices.

Fig. 4.9 represents the results from factory automation use case where the UEs send data only once based on an event. The event in the simulation is started by a random time from the start until 10 seconds of the simulation. Based on this random time value, all devices start sending data at that time. The latency remains below 10 ms with 2-os TTI, however, dynamic scheduling manages to keep the latency below 5 ms in this case as well. Therefore, narrowband UE categories can fulfill the URLLC latency requirements while trading-off the maximum number of devices in the cell.

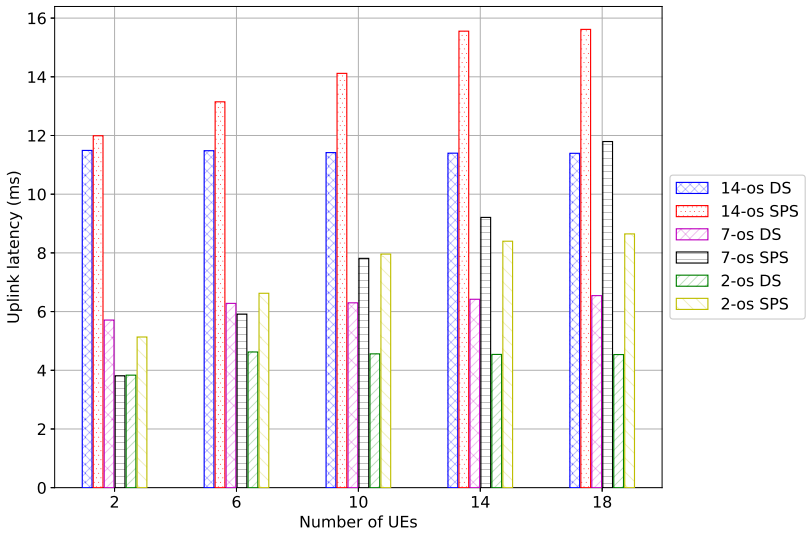

Figure 4.9: Uplink latency comparison of different TTI for industry automation use case with different number of devices where the LTE-M UEs send data only once based on some event such as a malfunction or an emergency stop. 2-os TTI outperforms legacy TTI with both type of scheduling. However, dynamic scheduling performs better in this case due to much sophisticated resource utilization.

A similar evaluation approach, as mentioned above for LTE-M, is taken for NB-IoT evaluation. The system bandwidth of NB-IoT is six times less than that of LTE-M (from 6 resource blocks to 1 resource block), therefore, the maximum number of evaluated NB-IoT UEs is ten. Figure 4.10 presents the simulations results form a factory automation scenario where 10 UEs send data to eNB.

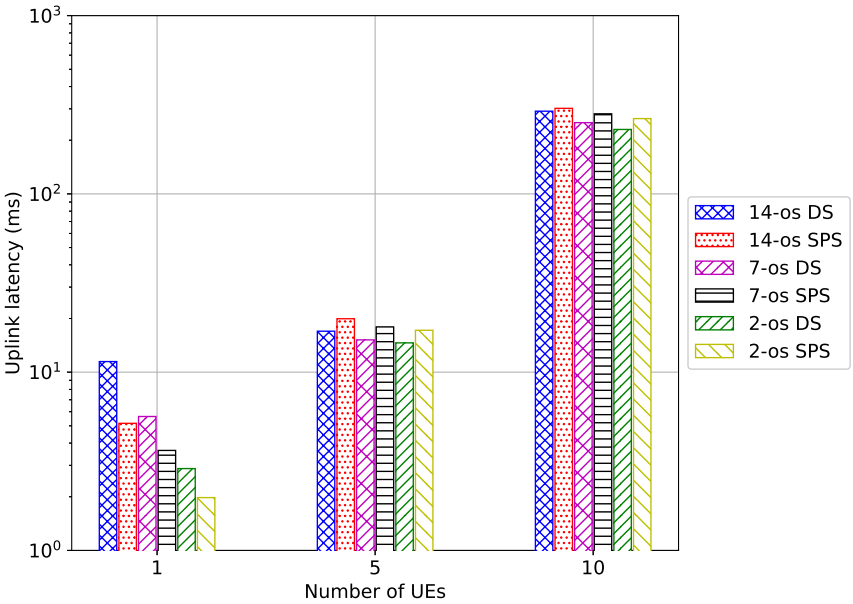

Figure 4.10: Uplink latency comparison of different TTI for factory automation use case with different number of NB-IoT UEs sending periodic data. 2-os TTI outperforms legacy TTI for one device with both type of scheduling. However, it is interesting to note that for a higher number of UEs, NB-IoT does not allow < 10 ms latency.

The eNB and UEs are placed inside a factory hall and hybrid building propagation loss model is used in the simulations for realistic evaluation. The hybrid building propagation loss model is a compound of different models able to evaluate the pathloss from 200 to 2600 MHz in different environments and with buildings (i.e. indoor and outdoor). The short TTI and semi-persistent scheduling perform well when the total amount of sent data traffic is below the system capacity. Here the total sent data traffic in case of five UEs is higher than the system capacity, therefore the uplink latency has a direct impact. NB-IoT is clearly not suitable for those URLLC applications that also require a higher data rate. However, it is interesting that NB-IoT can support very-low latency for those URLLC applications that have lower data rate requirements.

4.5.2.3 SPS and sTTI Evaluation for V2X

The simulation evaluation of sTTI for V2X use case where all the UEs are mobile (i.e. vehicles) was performed on the map of Offenburg city as shown in Figure 4.11 with Three Log-Distance propagation model and four eNBs installed at different locations in the city. The three log distance propagation loss model is based on the generic log distance propagation loss model with an addition of three different field, i.e. near, middle, and far, with different exponents. The evaluation is presented in Figure 4.12. The mobility traces of vehicles are generated with Simulation of Urban Mobility (SUMO) [108], which is a tool specific for this purpose. The vehicles are configured to move with a speed ranging between 30 kmph and 60 kmph. The mobility traces generated with SUMO are used in ns-3 as the mobility model for the UEs.

It can be noticed from the results in Fig. 4.12 that shorter TTI and SPS cannot fulfill 10 ms requirement for a comparatively larger number of UEs (i.e. >30). However, for smaller number of UEs, sTTI and SPS prove to be very effective in keeping the uplink latency under 10 ms (see Fig. 4.13). It is also important to mention here that reduction in TTI leads to an increase in control overhead, which obviously affects resource utilization. As compared to the industry automation use case simulation, the number of active eNBs is four times larger in V2X simulations due to the larger coverage area. This is also the reason behind very low difference in uplink latency for both use cases.

Fig. 4.13 represents the uplink latency for different number of UEs with periodic data uploads for V2X use case scenario. The overall simulation settings are similar to as in Fig. 4.12. The latency of legacy 14-os TTI with dynamic scheduling is always more than 10 ms, while shorter TTI (i.e. 2-os) with dynamic scheduling manages to keep the latency under 5 ms. SPS seems less effective from the result as the number of devices increase. It is due to the

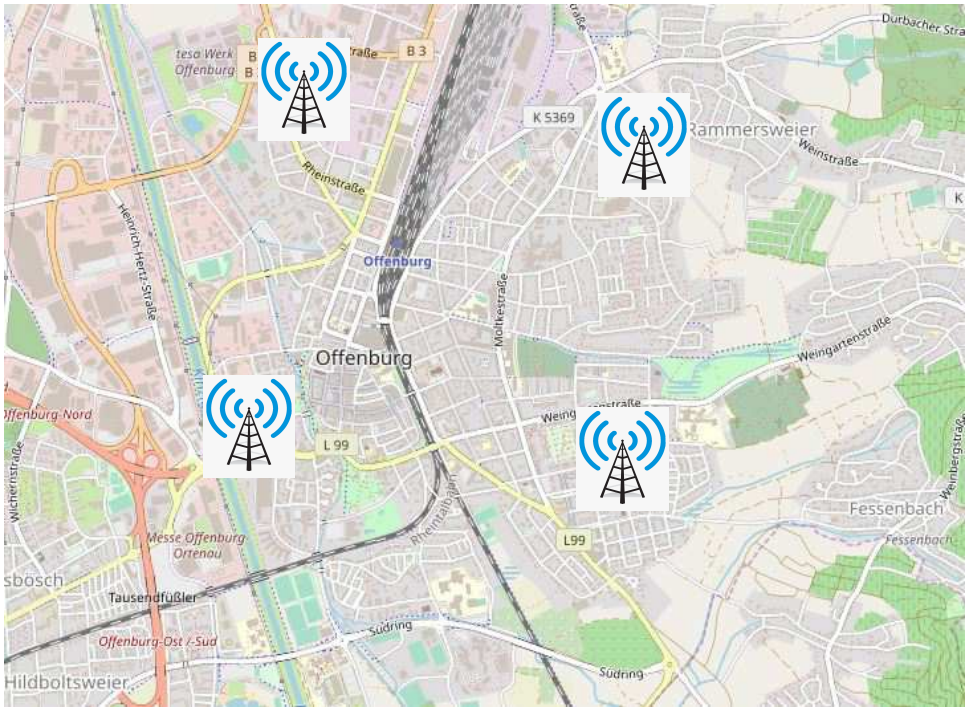

Figure 4.11: Map of Offenburg city used in the simulation.

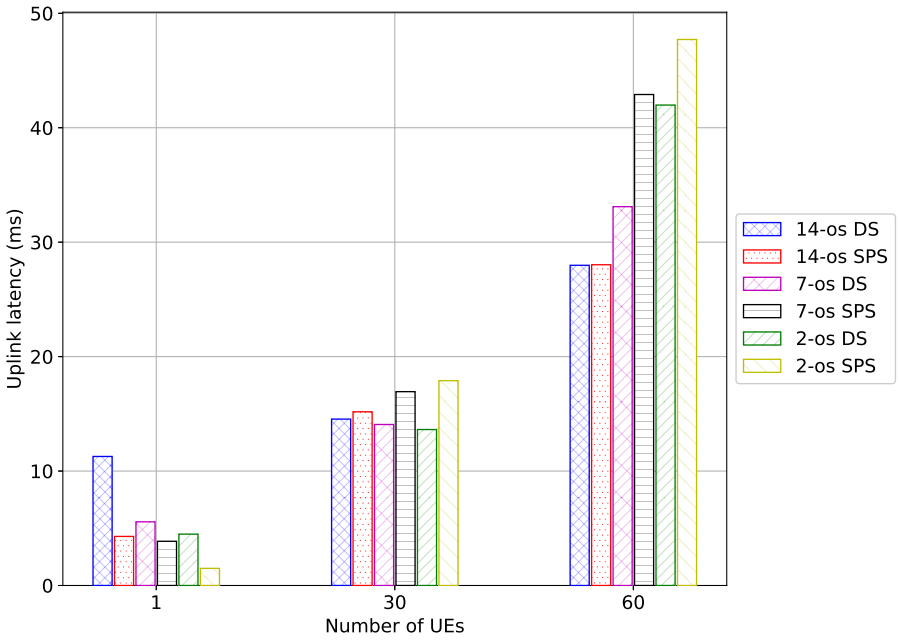

Figure 4.12: Uplink latency comparison of different TTI for V2X use case where multiple mobile LTE-M UEs send/receive data to/from four eNBs.

fact that with more devices, the periodicity of scheduled resources also increases. However, for very low number of devices, SPS outperforms DS in uplink latency.

The applications of V2X use cases require mobility of UEs. Since NB-IoT is not designed for such use cases with mobility and there are no hand-overs defined in the standards, V2X use cases is not evaluated for NB-IoT UEs in this thesis work. Moreover, the benefit of low-cost and

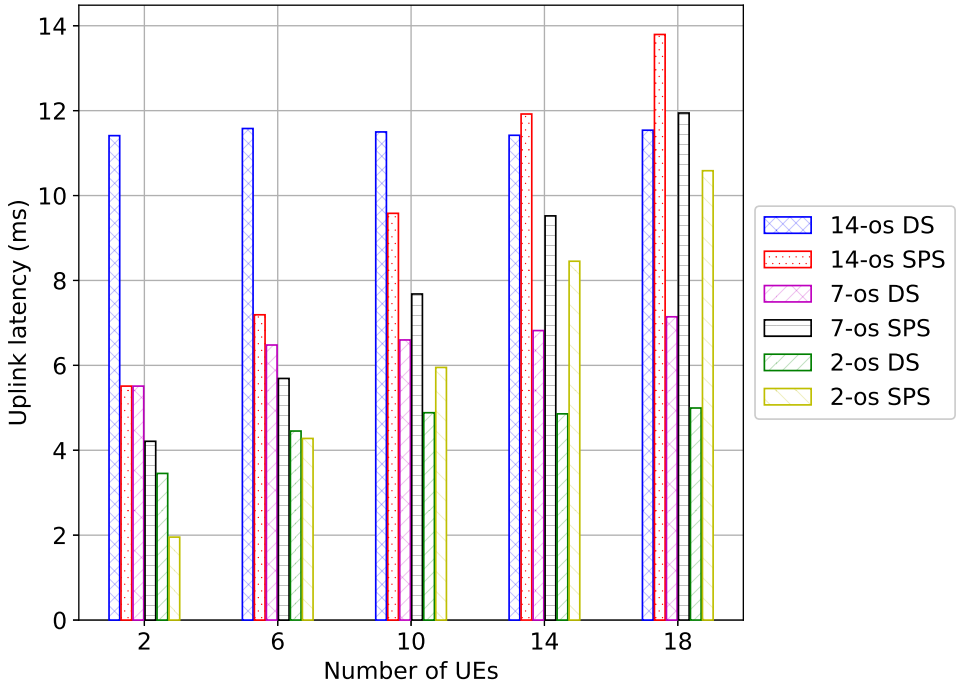

Figure 4.13: Uplink latency comparison of different TTI for V2X use case with different number of LTE-M UEs sending date periodically to eNB. 2-os TTI is clearly a better choice over legacy 14-os TTI. The short TTI can further reduce the latency to 2 ms with help of SPS but only for a very small number of devices.

low-energy consumption might not be as critical in V2X use cases as in industry automation. Therefore, the evaluation of NB-IoT for V2X is not considered.

4.6 Cost of sTTI and SPS

In LTE frame structure implementation of ns-3, in downlink, the first three symbols over entire bandwidth are reserved for control messages as shown in Fig. 4.14.

In uplink frames, the last symbol of each subframe is reserved for Sounding Reference Signal (SRS). According to ns-3 LTE design documentation [69], SRS allows of having every TTI an evaluation of the interference scenario since all the eNBs are transmitting (simultaneously) the control frame over the respective available bandwidths. The uplink design of 7-os TTI includes two transmissions of SRS within a subframe (one in each TTI). Similar to 7-os TTI, the SRS in 2-os TTI is sent multiple times within a subframe in every alternate TTI. Therefore, with 2-os TTI, the SRS is sent three times per subframe. The SRS occupies one symbol over the whole system bandwidth. In 14-os TTI, SRS occupies 7% of total system bandwidth, which becomes double (i.e. 14%) and triple (i.e. 21%) in 7-os and 2-os TTI respectively. In conventional LTE subframe design, the control overhead is roughly around 26% [107]. However, the LTE system with sTTI have extra control overhead of SRS, which can lead to a maximum of 40% of the

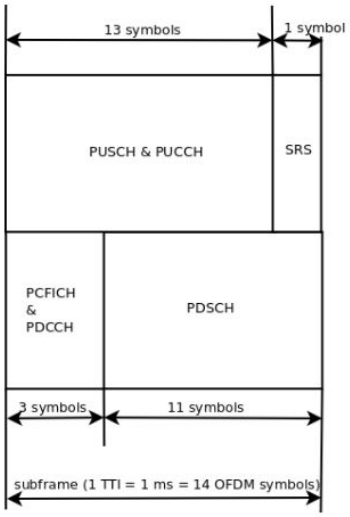

Figure 4.14: Subframe division in ns-3 LTE module according to [69]. First three symbols over entire bandwidth are reserved for downlink control signaling. In uplink, the last symbol is used as sounding reference signal.

total system bandwidth. Since all uplink control messages are an ideal messages and do not consume any radio resources, the effect of PUCCH overhead for sTTI cannot be observed from the simulation model. The reduction in latency presented in this chapter is achieved at the cost of increased control overhead from 27% to 40% (in case of 2-oss TTI).

The semi-persistent scheduler in ns-3 allocates network resources to UEs on a-priori basis. The maximum allocated resource at one scheduling instance is one RB. Let us take an example scenario where ten UEs are connected to the network and can send the data to eNB anytime. Assuming that one of the UEs needs to send an amount of data, which requires more than one RB, in dynamic scheduling, the scheduler allocates the required RBs to that UE so that the data could be sent once only if it is within the maximum allowed RB limit. However, in case of SPS, the scheduler only allocates one RB per UE on a round robin fashion, the UE with higher amount of data to send will transmit the date in multiple chunks of scheduled resources, which leads to an extra latency. Furthermore, if the other nine UEs did not have data to send in this time period, the allocated resources to those UEs by semi-persistent scheduler are wasted. This is the cost of latency reduction with SPS, which could be reduced by another approach proposed in this thesis in Chapter 5.

4.7 Conclusion

The short TTI and SPS can reduce uplink latency up to 85% at the cost of higher control overhead and resource utilization efficiency. Moreover, enabling support for multiple-sized TTI within the same network cell is very important and needs further evaluation. Obviously, the UEs that can only support legacy TTI cannot utilize sTTI, and therefore, providing backward compatibility for TTI is a necessary step towards enabling ultra-low latency 5G cellular networks.

In this chapter, the short TTI and SPS latency reduction techniques are described and the contributions from the literature for these techniques are listed. These techniques are evaluated to discuss their potential for supporting low latency use cases. The evaluation covered industrial automation applications with periodic traffic patterns as well as event-triggered traffic patterns, and V2X applications with periodic data uploads. The TTI lengths of 2-os, 7-os and legacy 14-os together with dynamic and semi-persistent scheduling are considered. The TTI length variants and SPS are implemented in the open-source ns-3 simulator and evaluated for narrowband LTE-M and NB-IoT UE categories. The results show that, for a single UE, short TTI of 2-os with SPS reduces the latency for more than 85% compared to the legacy TTI 14-os with dynamic scheduling. With an increased number of UEs, where some UEs send data periodically and others send data sporadically, the results show that 2-os short TTI with either SPS or DS can significantly reduce the latency. Thus, these combinations have the potential to support URLLC applications with stringent latency requirements.

5

Hybrid Scheduling & Flexible Transmission Time Interval

5.1 Résumé

L'ordonnancement le plus couramment utilisé dans le réseau LTE est l'ordonnancement dynamique (Dynamic Scheduling - DS) qui apporte une latence supplémentaire de quelques millisecondes. L'ordonnancement dynamique est conçu pour maximiser l'utilisation des ressources et pour fournir des débits de données maximums aux utilisateurs. D'autre part, l'ordonnancement semi persistant (Semi Persistent Scheduling – SPS) est conçu pour réduite les messages de contrôle de l'ordonnancement pour les transmissions de données périodiques. Cependant, SPS n'utilise pas efficacement les ressources du réseau en raison de l'attribution de ressources aux UE qui n'ont pas de données à transmettre. Les deux types de programmation ont leurs avantages et inconvénients. L'ordonnancement dynamique est conçu dans le but de maximiser le débit de données en utilisant efficacement les ressources du réseau. Les débits de données maximisés sont obtenus en incluant une procédure de demande de demande/octroi, qui nécessite quelques millisecondes de latence supplémentaire en liaison montante. SPS, quant à lui, est conçu pour minimiser les messages de demande/octroi de ressources pour les transmissions de données périodiques. En éliminant les messages de demande/d'octroi, SPS permet également d'obtenir une latence plus faible par rapport à l'ordonnancement dynamique, mais l'approche basée sur des messages de demande/d'octroi sans demande dans SPS entraîne un gaspillage des ressources du réseau, ce qui peut en fin de compte réduire les débits maximums de données.

Si les TTI réduits permet de réduire la latence, ils présentent certains inconvénients, notamment une augmentation des messages de contrôle et la prise en charge de plusieurs structures TTI dans une sous-trame. La surcharge de contrôle est augmentée par l'utilisation d'un TTI raccourci car la périodicité du signal de référence de sondage (Sounding Reference Signal – SRS) est réduite. Il réduit également les débits de données maximum du système en raison de la diminution des ressources disponibles. Un autre problème intéressant qui se pose avec l'implémentation du sTTI dans le simulateur ns-3 est la rétrocompatibilité. Comme les simulateurs ont des limitations sur différents paramètres, l'un des problèmes de la fonctionnalité sTTI dans ns-3 est que dans toute simulation, seulement un seul type de TTI peut être simulé. Cela peut poser un problème s'il est utilisé de la même manière en situation réelle. L'eNB qui prend en charge des TTI courts devrait également prendre en charge TTI standard de 14-os.

Afin de résoudre les problèmes d'ordonnancement LTE mentionnés ci-dessus, et de maximiser les performances en termes de latence pour les applications URLLC, une nouvelle approche d'ordonnancement est présentée dans ce chapitre. Elle combine les deux types d'ordonnancement et apporte des améliorations prometteuses dans les réseaux LTE 4G. L'ordonnanceur proposé s'appuie sur l'approche SPS de l'ordonnancement sans demande/octroi pour les UE de type URLLC, tout en gérant l'utilisation efficace des ressources de l'ordonnancement dynamique pour les UE tolérant la latence. L'ordonnanceur proposé est mis en œuvre et évalué dans le module LTE du simulateur ns-3.

Dans ce chapitre, une approche de TTI flexibles est présentée dans le but de réduire la question de l'augmentation messages de contrôle avec les TTI courts et de permettre la rétrocompatibilité des UE. Dans cette proposition, plusieurs longueurs de TTI sont prises compte dans la même largeur de bande du système et le même eNB. Cette proposition de TTI multiples a été mise en œuvre dans le module LTE du simulateur ns-3. Et évaluée afin comprendre les améliorations possibles que les multiples TTI peut apporter pour la prise en charge d'applications de type URLLC.

5.2 Introduction

Long Term Evolution (LTE) networks primarily focus on providing the high data rates for mobiles devices. The LTE networks were introduced so that they can meet the new requirements/challenges posed on the cellular network, which mainly aims at achieving greater performance and defining a new packet optimized architecture for the radio access network. As discussed in Chapter 4, the resource scheduling in LTE is managed by the eNB. The key idea of scheduling is to achieve a performance gain in the network in terms of higher data rate. The scheduler in LTE eNB performs many tasks, such as managing the queues, scheduling data packets, and transmitting those packets to the user nodes and allocating resources to these UEs. In LTE, scheduling is done at per subframe basis i.e. every 1 ms. To meet the strict latency requirements of URLLC use cases, among other approaches, sTTI and SPS have been proposed and investigated into this thesis (see Chapter 4). Two sTTI variants (i.e. 7-os and 2-os) along with legacy 14-os have been evaluated and compared. It is noticeable from the evaluation that sTTI and SPS can bring the latency of LTE network below 5 ms. Therefore, both techniques show a very promising potential to support URLLC applications.

The most commonly used scheduling in LTE network is dynamic scheduling, which unfortunately brings an additional uplink latency of a few milliseconds. The dynamic scheduling is designed to maximize the resource utilization and to provide maximum data rate to users. On the other hand, SPS is designed to minimize the control overhead of dynamic scheduling and is best fitted for periodic data transmission. However, as discussed in Section 4.3.2, SPS does not utilize network resources efficiently. Both of the scheduling types have their own advantages and disadvantages. Dynamic scheduling is designed with an aim to maximize data rates by efficiently utilizing the network resources. In turn the maximized data rates are achieved by including a resource request/grant procedure, which includes a few milliseconds of extra latency in uplink. SPS on the other hand is designed to minimize the scheduling request/grant messages for periodic data transmissions, that is particularly efficient for Voice over LTE (VoLTE) for example. By eliminating the request/grant messages, SPS also achieves low latency as compared with dynamic scheduling, however the request-less grant approach in SPS costs network resources being wasted, which can ultimately reduce the maximum data rates.

While sTTI brings the latency to very low values, it has a cost as mentioned in section 4.6, which includes an increased control overhead and support for multiple TTI structures within a subframe. The control overhead is increased by using a shortened TTI because the SRS periodicity is decreased. It also reduces the maximum system data rate due to fewer resources

available for the PUSCH. Another interesting problem that arrives with sTTI, specifically by the ns-3 simulator implementation is the backward compatibility. Since the simulators always have limitations on different parameters, one issue in ns-3 sTTI feature is that in any simulation, only one type of TTI can be simulated. This is a problem for the simulation of real world use cases that have to consider an eNB supporting sTTI as well as legacy 14-os TTI.

In order to resolve the above mentioned issues of LTE scheduling, and to maximize performance in terms of latency for URLLC applications, a novel scheduling approach is presented in this chapter, which combines both types of scheduling and provides promising improvements in 4G LTE networks. The proposed scheduler leverages from the schedule request/grant less approach of SPS for URLLC UEs while still manages the efficient resource utilization of dynamic scheduling for latency-tolerant UEs. The proposed scheduler is implemented and evaluated in LTE module of ns-3 simulator.

In this chapter, a flexible TTI approach is also presented with the aim to reduce the issue of increased control overhead with sTTI and support UE backward compatibility. In the proposed design, multiple lengths of TTI are supported within the same system bandwidth and eNB. The proposed TTI design is implemented in ns-3 simulator within the LTE module. Afterwards, the developed feature is evaluated through simulations to understand the possible improvements that flexible TTI brings in supporting URLLC applications.

In order to explain the design of hybrid scheduler, the common LTE schedulers are discussed in the following. The design and evaluation of hybrid scheduler are then presented afterwards.

5.3 Scheduling in LTE

The scheduling process is used by eNB to decide which UE will have transmitting and receiving resources and by how much. In LTE, scheduling is done at per subframe basis i.e. every 1 ms. The entity, which governs the resource utilization and allocation in eNB is called a scheduler. In the following, two most types of scheduling being used in LTE systems are discussed.

5.3.1 Dynamic Scheduling

Dynamic scheduling is based on request/grant principle where UEs send a scheduling request when they require resources to send or receive data. In response to that request, eNB sends a scheduling grant mentioning the number of resource blocks in the LTE time frequency resource grid. Every transmission time interval, MAC layer of eNB checks for the UEs to be scheduled, the data availability for each UE to be scheduled, and the feedback from the UE on the channel conditions. Based on these data, eNB schedules the resources for the UE through the PDCCH. If data is not available, UE does not get scheduled resources.

The dynamic scheduling aims to maximize the data rates by efficiently utilizing the network resources. However, it is achieved at the expense of control overhead of scheduling request/grant. As depicted in Figure 3.1, after the data is available at UE to be sent out to eNB, a request must be sent to acquire the resources for data transmission. The cost of this approach is the extra latency of around 6 ms in uplink as mentioned in Table 3.1. In a context where the target is a latency of less than 10 ms, this is a bottleneck problem. The maximized data rate is achieved at the expense of extra latency.

5.3.2 Semi-Persistent Scheduling

Semi-persistent scheduling was originally designed for applications such as VoLTE that require periodic data transmissions and the periodicity of the transmissions in known in advance. Therefore, the resource are allocated to the UEs by eNB on a-priori basis. The scheduling is in SPS is independent of the data available at the UEs. In other words, if there is no data to send at UE, it will still be scheduled after a certain time interval. Figure 4.3 shown the message flow of SPS. The eNB after configuring the SPS in uplink and downlink, allocates resources to UEs and sends this information in scheduling grant message. As soon as the UEs have data available to send to eNB, they use the respective scheduled resources to send it. It is important to note that by eliminating resource request message, SPS is able to reduce around 6 ms from the uplink (see Table 3.1).

The LTE network normally utilizes dynamic scheduling for high data rate applications. SPS is only used for VoLTE since it was proposed for the purpose of supporting applications that need continuous uplink resources. However, in Release 13, 3GPP proposed that SPS can also be used for latency reduction in LTE networks. The cost of using SPS is the wastage of resource, which results from the UEs that do not have data to send in the scheduled resources. This phenomenon reduces the maximum data rate.

5.4 Hybrid Scheduling for LTE

The hybrid scheduling is developed as part of this thesis work, which basically uses the dynamic and semi-persistent scheduling within a single eNB [1]. Depending upon the requirements of the UE, eNB can schedule resources for UE on either on demand or a-priori basis. This technique helps in minimizing the cost of both the scheduling types. The eNB segregates UEs based on their types i.e. latency-critical UE or latency-sensitive UE and schedules the resources accordingly. In order to clearly present the design of hybrid scheduler and its implementation, we first present the working mechanism of resource scheduling in LTE module of ns-3 in the following.

5.4.1 LTE Scheduling in ns-3

There are different types of dynamic schedulers implemented in ns-3 simulator. A scheduler is in charge of generating specific structures called Data Control Indication (DCI), which are then transmitted by the PHY of the eNB to the connected UEs, in order to inform them about the resource allocation on a per subframe basis. In doing this in the downlink direction, the scheduler has to fill some specific fields of the DCI structure with information, such as: the Modulation and Coding Scheme (MCS) to be used, the MAC Transport Block (TB) size, and the allocation bitmap, which identifies the RBs that will contain the data transmitted by the eNB to each user [69].

For the mapping of resources to physical RBs, ns-3 adopts a localized mapping approach, hence in a given subframe each RB is always allocated to the same user in both slots. The allocation bitmap can be coded in different formats; in this implementation, the Allocation Type 0 is considered, according to which the RBs are grouped in Resource Block Groups (RBG) of different size determined as a function of the transmission bandwidth configuration in use.

[1] The hybrid scheduling and its implementation design for ns-3 was conceptualized by the author and implemented by S. R. Gawda as part of his Masters degree internship.

For certain bandwidth values not all the RBs are usable, since the group size is not a common divisor of the group. This is for instance the case when the bandwidth is equal to 25 RBs, which results in a RBG size of 2 RBs, and therefore 1 RB will result not addressable. In uplink the format of the DCIs is different, since only adjacent RBs can be used because of the Single Carrier Frequency Division Multiple Access (SC-FDMA) modulation. As a consequence, all RBs can be allocated by the eNB regardless of the bandwidth configuration.

The already implemented LTE schedulers available with the latest version of ns-3 are all based on dynamic scheduling [106]. The semi-persistent scheduling was developed as part of the work in this thesis. More details are given in section 4.3.2 of this thesis. In the following, the design and implementation of hybrid scheduling is discussed in more detail.

5.4.2 Hybrid Scheduler Design

Every simulation in ns-3 is run using a configuration script, which defines the node positions, mobility, protocol stack, and the traffic model for communication between them. For LTE simulations, ns-3 allows multiple eNBs in a network with multiple cells within one eNB by using directional antennas. The application running on top of LTE protocol stack can also be configured depending on the simulation scenarios. However, the MAC layer of LTE do not support the use of multiple schedulers within a single simulation. The scheduler can be configured by the *SetSchedulerType* function of *LteHelper* class, which configures a given scheduler for all the eNBs in the simulation. Therefore, within a single simulation run, there is no possibility in ns-3 to use two different schedulers or to assign a different scheduler to eNBs.

In *Hybrid Scheduling* implementation, the UEs are divided into two groups, latency-sensitive UEs and latency-tolerant UEs. The rationale behind dividing the UEs into two groups is that it is more realistic to consider the presence of both type of UEs in the real world scenarios. Each UE is assigned a latency-sensitive flag for the scheduler to know if it requires low latency. This information is then utilized by the eNB to schedule the resources accordingly. Latency-tolerant UEs are scheduled by the DS to maximize throughput, while latency-sensitive UEs are scheduled by SPS to minimize latency. Figure 5.1 represents the working mechanism of hybrid scheduling. The eNB schedules latency-tolerant UEs (G1) with dynamic scheduling while latency-sensitive UEs (G2) are scheduled through semi-persistent scheduling. The hybrid scheduler uses round robin implementation from ns-3 LTE module and SPS from the in-house developed scheduler (see Section 4.3.2).

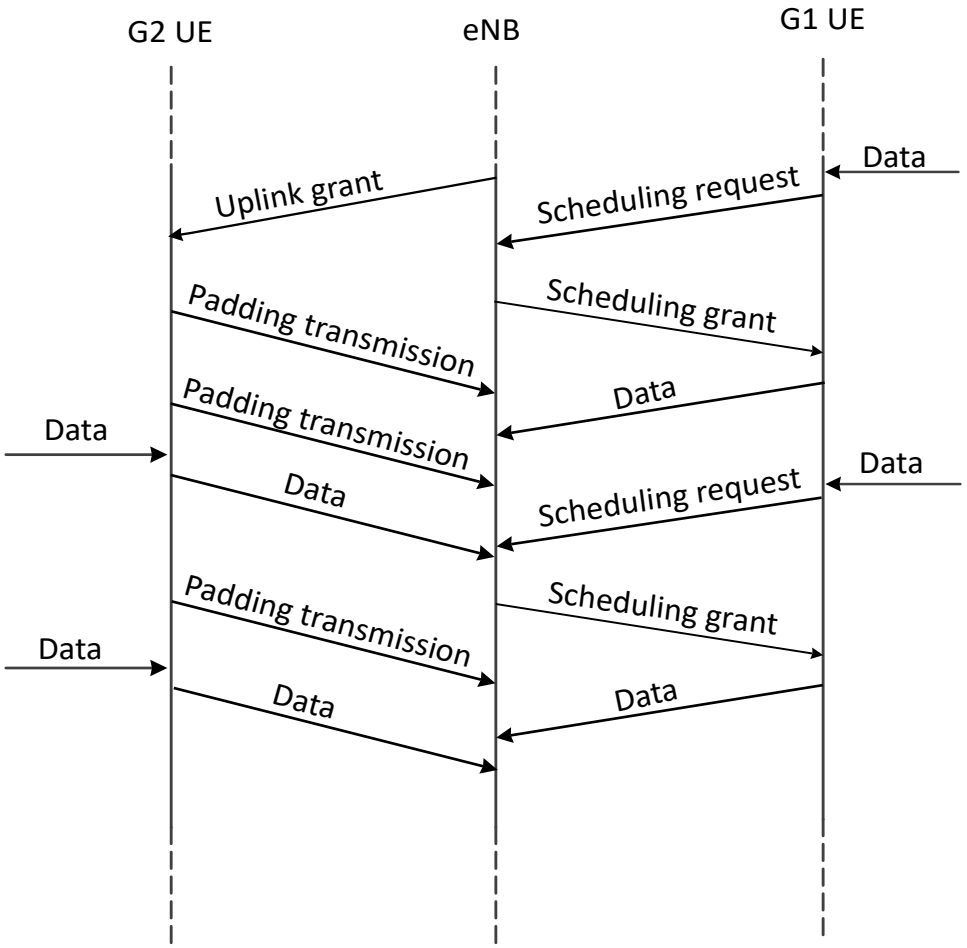

Figure 5.1: Hybrid scheduling working mechanism: UEs belong to either one of two groups. Group 1 (G1) represents latency-tolerant UEs that use dynamic scheduling while Group 2 (G2) represents latency-sensitive UEs that use semi-persistent scheduling. The eNB schedules resources according to the group of UEs with either dynamic scheduling or semi-persistent scheduling.

5.4.3 Experimental Evaluation of Hybrid Scheduling

The hybrid scheduler developed for ns-3 in this thesis is evaluated over different scenarios in this section. Following subsections describe overall simulation setup and results obtained from simulations.

5.4.3.1 Simulation Setup

The parameters used in the simulations for this section are shown in Table 5.1. The evaluation of hybrid scheduling approach is performed for LTE-M UE category with 1.4 MHz bandwidth. In order to analyze the potential of hybrid scheduler, the following four scenarios are simulated.

The UEs in the simulation are divided into two groups named G1 and G2. In all the simulations, UEs in G1 group are considered latency-tolerant and use dynamic scheduling, while UEs in G2 are considered as latency-sensitive and use semi-persistent scheduling. The evaluated scenarios are listed in the following.

Table 5.1: Simulation parameters used in ns-3 simulations for the evaluation of uplink latency with proposed novel hybrid scheduling approach.

Parameter	Value
Simulator - version	ns - 3.26
eNB transmission power	43 dBm
UE transmission power	20 dBm
Uplink bandwidth	1.4 MHz
Number of resource blocks	6
Packet size	50 B
Packet transmission interval	100 ms
Transmission time interval (TTI)	1 ms
Resource scheduling	Hybrid
Simulation duration	30 s

a. Two UEs in a cell where one belongs to G1 and other to G2. The simulation scenario evaluates dynamic, semi-persistent, and hybrid scheduling to compare the performance in terms of latency.

b. The performance evaluation of hybrid scheduling for different number of UEs in the network. In this simulation scenario, G2 UEs are 10% of total UEs and remaining are G1 UEs.

c. The performance evaluation of hybrid scheduling for different number of UEs in the network. In this simulation scenario, G2 UEs are 20% of total UEs and remaining are G1 UEs.

d. The performance evaluation of hybrid scheduling for different percentage of latency-sensitive UEs in the network. In this simulation, total number of UEs is fixed to 30.

5.4.3.2 Results and Analysis

Case a Figure 5.2 shows the uplink latency from a simulation of 2 UEs in a network. In this simulation, one UE is considered as latency-tolerant (i.e. it uses DS) and other as latency-sensitive (i.e. it uses SPS). The results are for all three type of scheduling techniques, i.e. dynamic, semi-persistent, and hybrid. The uplink latency in case of dynamic or semi-persistent scheduling remains the same for both UEs in the simulation. However, with hybrid scheduling, it can be noted that the latency-sensitive UE is scheduled on semi-persistent basis and latency-tolerant UE receives resource on demand. Thus, the uplink latency of latency-sensitive UE is below 4 ms. The result shows that an intelligent blend of different scheduling techniques can prove to be very effective for different applications.

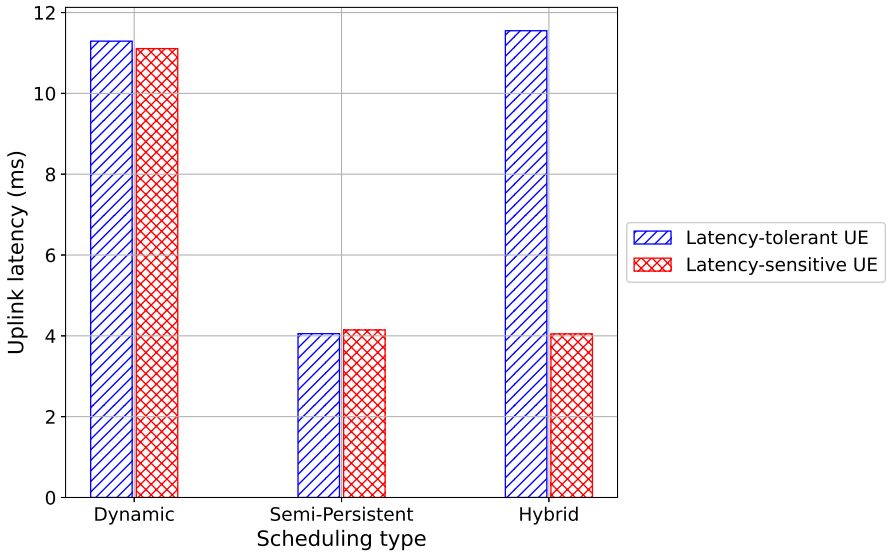

Figure 5.2: Scenario 1: Comparison of dynamic, semi-persistent, and hybrid scheduling for two UEs. In hybrid scheduling, latency-tolerant UE is scheduled with dynamic scheduling while latency-sensitive UE is scheduled with semi-persistent scheduling.

Case b Figure 5.3 shows the evaluation of hybrid scheduling for different device densities. The latency-sensitive UEs (i.e. G2) are 10% of total UEs, whereas the latency-tolerant UEs (G1) are the remaining 90%. All UEs in the simulation send data periodically to the eNB.

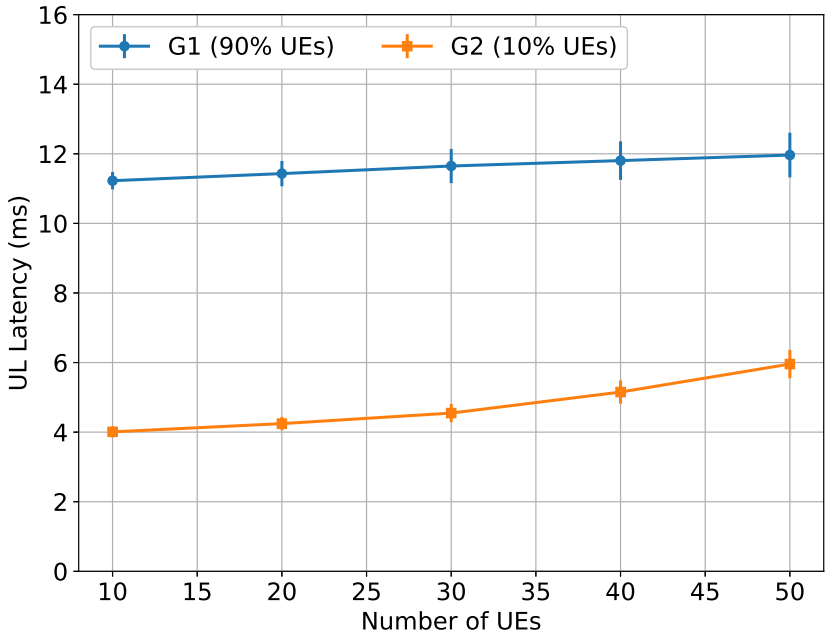

Figure 5.3: Scenario 2: Evaluation of hybrid scheduling for different number of UEs. G2 UEs are 10% of total UEs and remaining are G1 UEs. The latency-sensitive G2 UEs are always scheduled on a-priori basis and therefore achieve latency below 5 ms.

It is obvious from the result that the hybrid scheduling can successfully achieve below 10 ms latency for a 10% latency-sensitive UEs even for a higher device density (i.e. 50 UEs in a network cell). Moreover, the uplink latency is kept under 5 ms for lower device densities (< 40). Such low latency is achieved by allocating the network resources without demand to G2 UEs to guarantee that the latency requirement is met.

Case c Figure 5.4 represent a similar scenario as Figure 5.3. The only difference in Figure 5.4 is that G2 UEs are 20% and G1 UEs are 80% of total UEs. This particular scenario is selected to evaluate a higher number of latency-sensitive UEs in the network. Again, the hybrid scheduler keeps the uplink latency below 10 ms for G2 UEs in high device density (i.e. 50 UEs). The reason for a higher uplink latency as compared to Figure 5.3 is the network resources that are reserved for G2 UEs and being shared among all connected G2 UEs. In case of 20% G2 UEs, same amount of network resources is shared among UEs, which results in a slightly higher uplink latency.

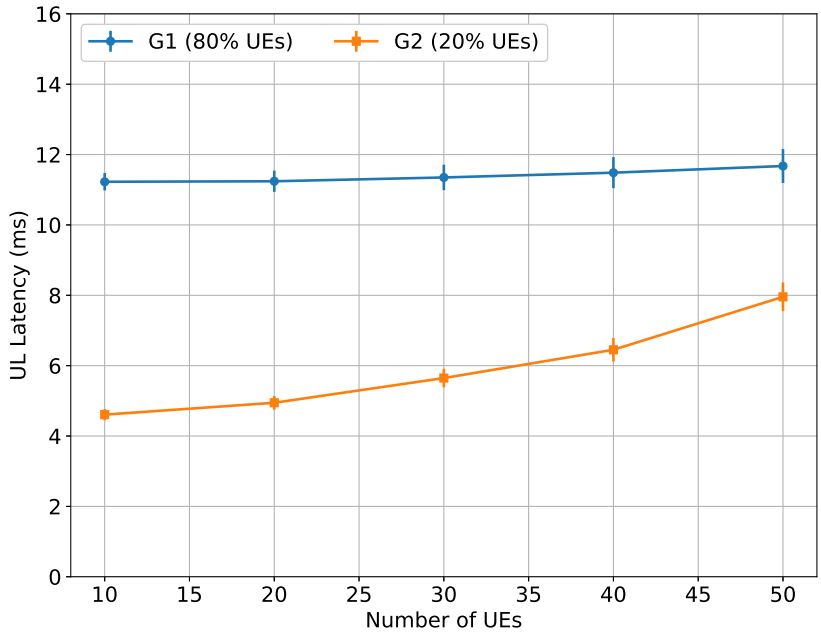

Figure 5.4: Scenario 3: Evaluation of hybrid scheduling for different number of UEs. G2 UEs are 20% of total UEs and remaining are G1 UEs. The latency-sensitive G2 UEs are always scheduled on a-priori basis and therefore achieve latency below 10 ms.

Case d Another interesting scenario is presented in Fig. 5.5 where 30 UEs in the network are simulated with hybrid scheduling. The percentage of latency-sensitive UEs is varied from 10% to 90% of total UEs. The uplink latency for 10% G2 UEs is below 5 ms, whereas in case of 90% G2 UEs, latency is four times increased. The reason behind this increase is the fact that

with more latency-sensitive UEs in the network, the resources are continuously allocated to UEs without any request, which leads to an increased latency. The latency for latency-tolerant UEs remains almost constant for different number of UEs since the resources are allocated on a request/grant procedure with dynamic scheduling. It is also interesting to note from the results that for 30 UEs in a network, hybrid scheduling keeps the uplink latency below 10 ms for almost 45% of G2 UEs. This also indicates that hybrid scheduling can successfully meet low latency requirements for a reasonable devise density.

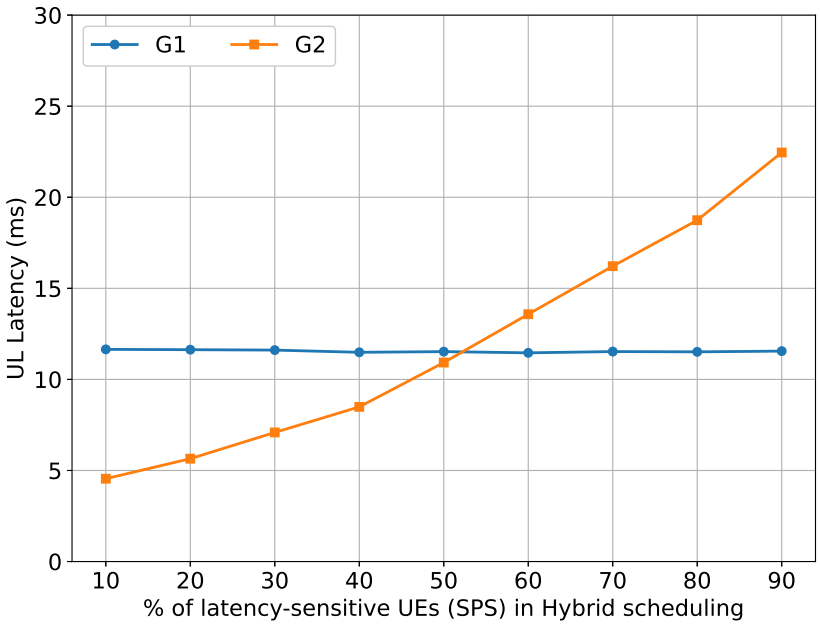

Figure 5.5: Scenario 4: Evaluation of hybrid scheduling for 30 UEs in a network with a variable percentage of devices among them being as latency-sensitive UEs. The x-axis represents the percentage of latency-sensitive UEs in the network. Two lines represent the uplink latency of G1 (UEs scheduled with DS) and G2 (UEs scheduled with SPS). For example, at 50% latency-sensitive UEs, half UEs in the network are scheduled with dynamic while other half is scheduled with semi-persistent scheduling.

5.5 Flexible Transmission Time Interval

The flexible TTI approach presented in this chapter is based on the sTTI techniques (discussed in Chapter 4). LTE system defines 10 ms radio frame to conduct the signal transmission. A radio frame contains ten 1 ms subframes, which are also called TTI as shown in Fig. 5.6. Shortened TTI utilizes time less than 1 ms as the transmission unit. With the reduction of the legacy TTI length of 1 ms, i.e. 14-os, the overall data transmission and processing time can be reduced. This reduces the latency for all data transmissions. The initial implementation of sTTI, as discussed in Chapter 4, was designed with the goal to evaluate different lengths of TTI. However, it was noticed from the evaluations that the support of legacy TTI alongside sTTI is also required due to fact that even with sTTI deployed and supported by eNB, there will still be some latency-tolerant UEs in the network that would operate only with legacy TTI. To this end, flexible TTI structure was developed and evaluated. The design, implementation and evaluation are presented in the following.

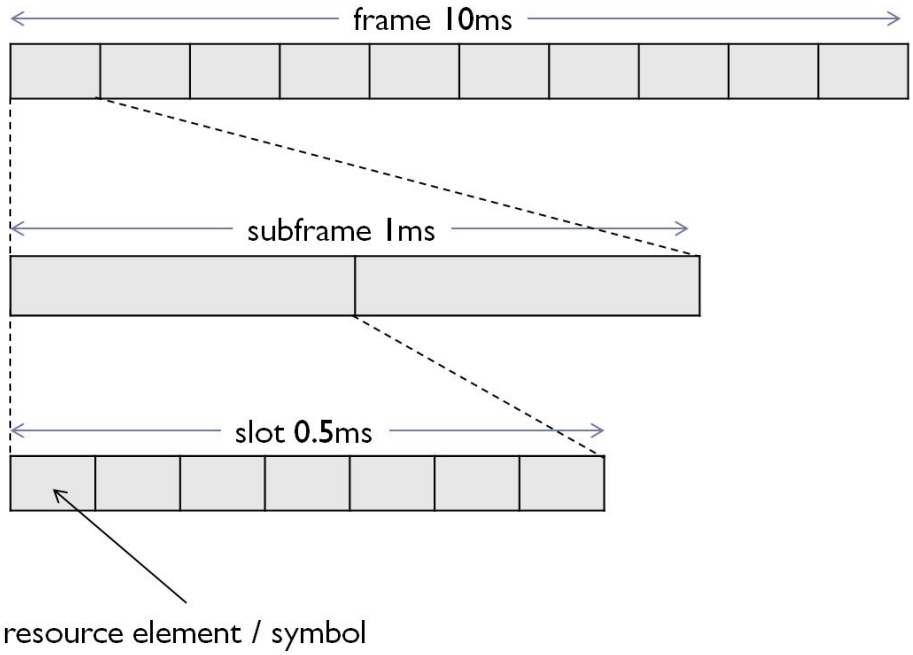

Figure 5.6: LTE frame structure: one frame consists of ten 1 ms subframes. Each subframe has 14 symbols/resource elements. Transmission time interval in legacy LTE is based on one subframe i.e. 1 ms.

5.5.1 Flexible TTI implementation in ns-3

The LTE module in ns-3 simulator implements each layer of protocol stack in form of blocks. Some of the layers even are implemented in multiple blocks, for example, the resource schedulers are not implemented as part of the MAC layer, rather they are separate blocks. Since the TTI is mainly a function of PHY layer, all developments were made initially in the PHY block of

the LTE module. However, the flexible TTI requires not only a change in PHY block, but also in the scheduler. The sTTI design is shown in Fig. 5.7. The network resources are sliced for different TTI groups. UEs in a specific group are only allowed to send/receive in their respective resource group. The implementation steps are as following.

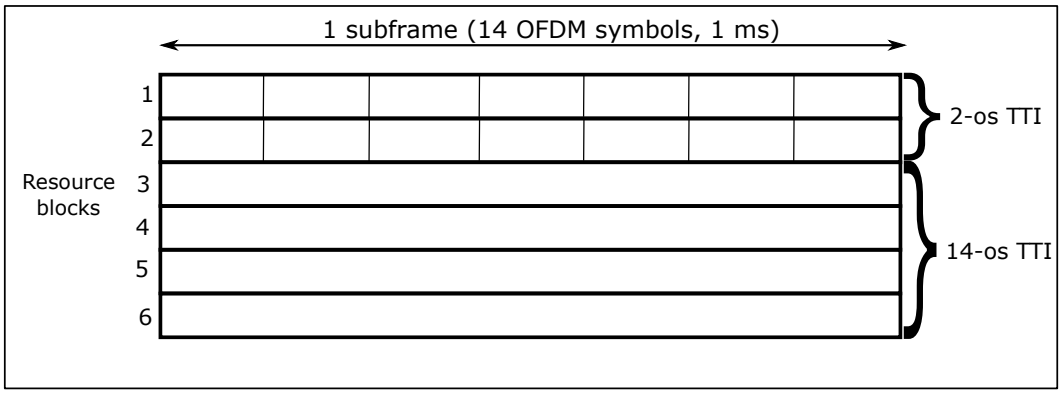

Figure 5.7: Flexible TTI design where two out of six resource blocks are used for 2-os TTI and remaining for legacy 14-os TTI. The resource management is done by the eNB.

- The TTI is specified by the *lte-phy* class of LTE module. This parameter is used by the *lte-spectrum* class where the actual transmissions are made. In the current open source release of ns-3 LTE module, TTI and subframe length are determined by only a single variable since they are same in legacy LTE. In sTTI implementation, these two parameters are made independent to each other. The subframe indication is performed in *lte-mac* class while TTI is determined in *lte-phy* class of the LTE module. On the UE side of protocol stack, the TTI value is fixed to either 14-os, 7-os, or 2-os. On the eNB side, in the *lte-phy* class, a list of UE identifier and TTI being used for that specific UE is maintained. With this list, the eNB makes transmissions in lengths according to the TTI of the UEs addressed.

- The system bandwidth is sliced into multiple slices. These slices represent resource blocks. For each supported TTI, a number of RBs are reserved. For example, with a system bandwidth of six RBs and three TTI variants, two RBs are reserved for each TTI. Which means, the first two RBs are used for 14-os, next two for 7-os, and last two for 2-os TTI. The control channels that occupy multiple RBs across the system bandwidth, remain the same. Only the shared channel for data in uplink and downlink are sliced in this way. This slicing approach is implemented in the *ff-mac-scheduler* class, which provides an abstraction for all the schedulers implemented in LTE module. Here the major changes are on the eNB side as most of the resource management is done by eNB. UEs are only allocated the resource blocks according to their TTI class.

The current implementation of flexible TTI is based on only round robin scheduler for the evaluation purpose, however, it can easily be adopted for the other schedulers available in ns-3 LTE module. In the following, the implemented features are validated through simulations.

5.5.2 Experimental Evaluation of Flexible TTI

The flexible TTI feature developed in ns-3 is evaluated with different scenarios in this section. The aim of this evaluation is to validate the proposed flexible TTI feature for a varying number of UEs in the network. Following subsections describe overall simulation setup and results obtained from simulations.

5.5.2.1 Simulation Setup

The parameters used in the simulations for this section are shown in Table 5.2. The evaluation of flexible TTI is performed for LTE-M UE category with 1.4 MHz bandwidth. For the evaluation, the following four scenarios are simulated.

Table 5.2: Simulation parameters used in ns-3 simulations for the evaluation of uplink latency with proposed flexible TTI approach.

Parameter	Value
Simulator - version	ns - 3.26
eNB transmission power	43 dBm
UE transmission power	20 dBm
Uplink bandwidth	1.4 MHz
Number of resource blocks	6
Packet size	50 B
Packet transmission interval	100 ms
Transmission time interval	1 ms, 0.14 ms
Resource scheduling	Dynamic
Simulation duration	30 s

The UEs in the simulation are divided into two groups named G1 and G2. In all the simulations, UEs in G1 group are considered latency-tolerant and UEs in G2 are considered as latency-sensitive.

- Two UEs in a cell where one belongs to G1 and other to G2. The simulation scenario compares the performance of legacy 14-os TTI and flexible TTI where flexible TTI uses a combination of 14-os and 2-os TTI.

- Different number of UEs in the network where G2 UEs are 10% of total UEs and remaining are G1 UEs.
- Different number of UEs in the network where G2 UEs are 20% of total UEs and remaining are G1 UEs.
- Different percentage of latency-sensitive UEs in the network where total number of UEs is fixed to 30.

5.5.2.2 Results and Analysis

A basic evaluation of flexible TTI with only two UEs in the network is shown in Fig. 5.8. Among both UEs, one is considered to be latency-sensitive and other as latency-tolerant. With both UEs using 14-os legacy TTI, the uplink latency is around 11.5 ms for both UEs. With flexible TTI, the latency-tolerant UE uses 14-os TTI for transmissions and latency-sensitive UE uses 2-os TTI. The data sent by latency-sensitive UE with flexible TTI has a latency of less than 3 ms. As explained earlier, the eNB schedules latency-sensitive UEs in the RBs reserved for short TTI UEs. In this simulation, two out of six RBs are for sTTI and four for legacy TTI. It is clear from the result, that flexible TTI approach can achieve very low latency while still supporting the legacy TTI UEs.

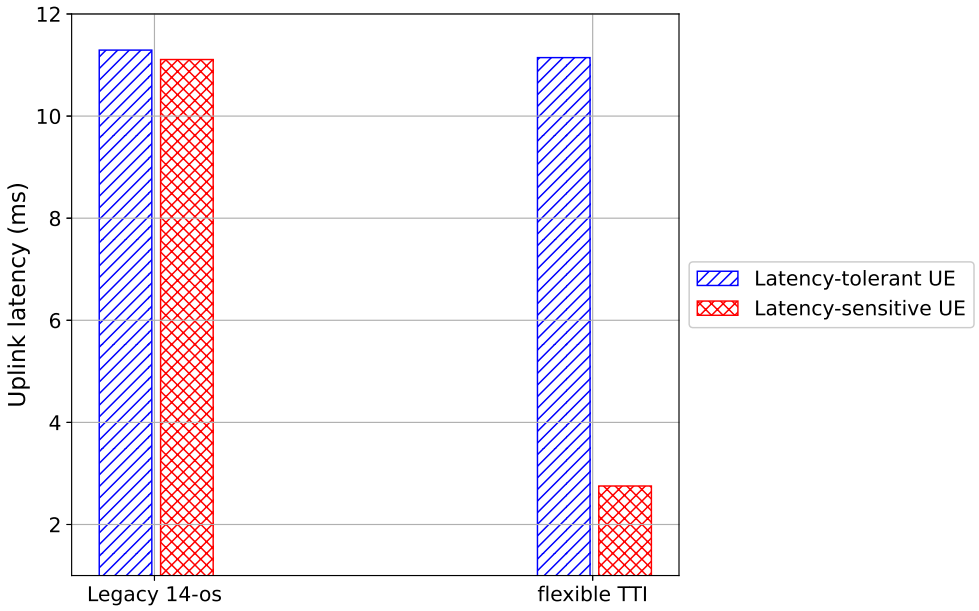

Figure 5.8: Scenario 1: Comparison of flexible TTI approach with legacy 14-os TTI. On the left side, with 14-os TTI, both UEs have almost similar latency of around 11.5 ms. However, in case of flexible TTI, the latency-sensitive UE has a latency of around 3 ms, which is achieved by using 2-os TTI.

An extended analysis of flexible TTI is presented in Fig. 5.9 and 5.10 with a higher number of UEs in the network. Figure 5.9 shows an evaluation of different number of device where latency-sensitive (G2) UEs are 10% of total UEs. All UEs in the simulation send data periodically to the eNB. Flexible TTI approach allows different TTI supporting UEs within the same eNB. It is noted from the results that with flexible TTI, the latency-sensitive UEs using 2-os TTI have a delay of less than 4 ms even when there are 50 total UEs in the network.

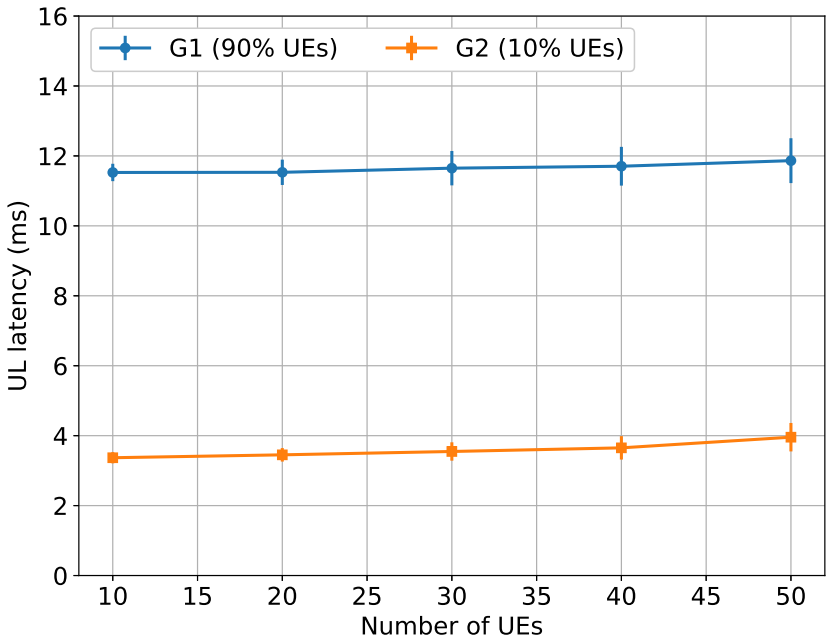

Figure 5.9: Scenario 2: Evaluation of flexible TTI for an increasing number of UEs. G2 UEs are 10% of total UEs and remaining are G1 UEs. The latency-sensitive G2 UEs use 2-os TTI for transmissions and therefore achieve latency below 4 ms.

Figure 5.10 also shows similar evaluation but with 20% G2 UEs in the network. In this case, the uplink latency is slightly higher. However, in both simulation scenarios, flexible TTI successfully achieves low latency for latency-sensitive UEs due to fact that these UEs use sTTI for uplink data transmissions.

To further extend the analysis, Fig. 5.11 presents an evaluation of 30 UEs in the network with a varying percentage of latency-sensitive UEs from 10% to 90% of total UEs. The uplink latency remains below 10 ms for all number of G2 UEs that use 2-os TTI. The increase in the uplink latency of G2 UEs is due to the fact that only two out of six RBs are used by eNB to schedule G2 UEs. As the number of UEs increase, the latency also increases due to limited resources. It is interesting to note from these results that with flexible TTI, the LTE network can support both latency-sensitive UEs and latency-tolerant UEs. Furthermore, the flexible TTI also supports legacy UEs that use 14-os TTI for transmissions.

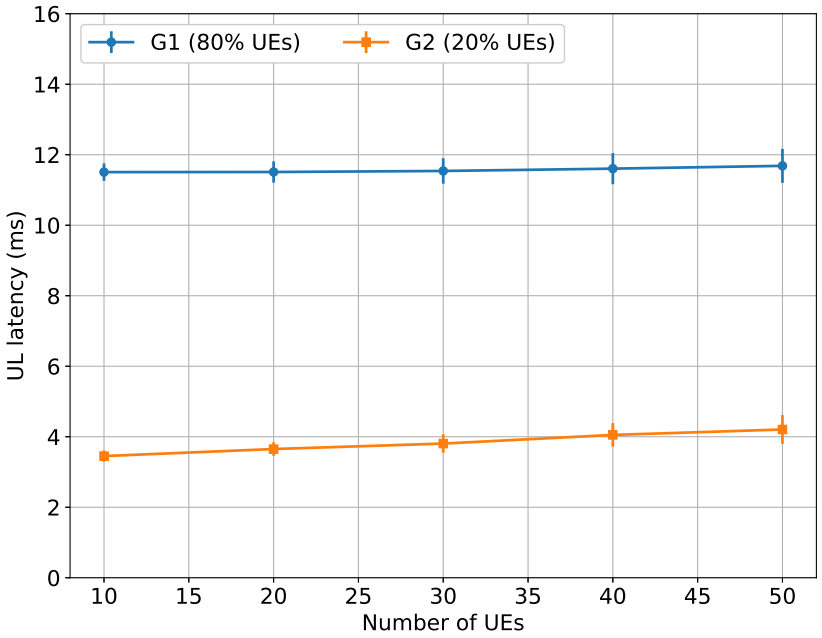

Figure 5.10: Scenario 3: Evaluation of flexible TTI for an increasing number of UEs. G2 UEs are 20% of total UEs and remaining are G1 UEs. The latency-sensitive G2 UEs use 2-os TTI for transmissions and therefore achieve latency around 4 ms.

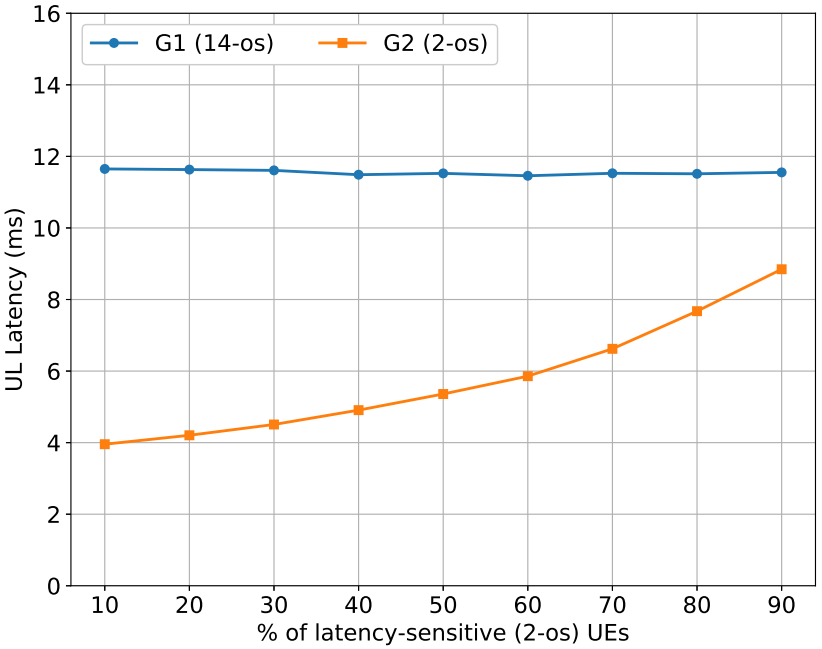

Figure 5.11: Scenario 4: Evaluation of flexible TTI for 30 UEs in a network with a variable percentage of devices among them being as latency-sensitive UEs. The x-axis represents the percentage of latency-sensitive UEs in the network. Two lines represent the uplink latency of UEs using 14-os and 2-os TTI. For example, at 50% latency-sensitive UEs, half UEs in the network use 14-os TTI while other half use 2-os TTI.

5.6 Conclusion

The novel use cases of MTC require different conditions to be fulfilled by the cellular networks. The network is not only expected to provide low-latency but also to support a large number of UEs that have different requirements concerning their latency. It has become essential to discover new techniques that can enable the network to meet the requirements. Latency reduction techniques presented and evaluated in Chapter 4 show a great potential in reducing the latency of 4G cellular networks. However, there are certain costs of these techniques that increase either the control overhead, or the network resource wastage.

In this chapter, two original approaches for resource scheduling and frame structure called hybrid scheduling and flexible TTI respectively are proposed and evaluated to reduce the cost of SPS and sTTI. The former makes use of dynamic and semi-persistent scheduling at the same time while fulfilling the latency requirements of URLLC use cases. While the later encapsulates multiple sized TTI within LTE subframes to support UEs with different latency requirements. Both approaches were implemented in ns-3 LTE module and evaluated over multiple scenarios to identify the advantages and disadvantages of using them. It has been noted that hybrid scheduling outperforms the traditional dynamic or semi-persistent scheduling for periodic data transmissions. The results also exhibit a promising potential in flexible TTI approach for allowing low-latency support for different applications as part of 5G cellular networks.

6

Conclusion and Outlook

6.1 Résumé

L'une des exigences les plus récentes en matière de réseaux cellulaires est de permettre la prise en charge d'applications IdO critiques. La demande croissante de ces applications en matière de latence pose un défi aux réseaux cellulaires. Pour répondre à des exigences de faible latence, de multiples approches ont été proposées et discutées dans un passé récent. Dans cette thèse, un travail de réduction de la latence dans les réseaux cellulaires 4G et 5G a été réalisé en proposant, développant et évaluant des techniques multiples de réduction de la latence du plan utilisateur. Plus précisément, les questions de recherche mentionnées au Chapitre 1 ont trouvé une réponse dans cette thèse.

Il y a cependant encore quelques directions qui pourraient être étudiées dans le cadre de travaux futurs à cette thèse :

- Intégration des techniques de réduction de la latence (issues de cette thèse) dans le cœur des réseaux de la prochaine génération afin d'analyser et d'étudier la latence globale de bout en bout des réseaux cellulaires. Comme les délais du cœur de réseau et de l'Internet varient, il serait très intéressant d'inclure tous les blocs de l'ensemble du système dans des travaux d'amélioration de la latence.

- L'une des applications de la 5G est les communications massives de machine à machine (Massive Type Communication – MTC) où le réseau devrait supporter une très grande densité de dispositifs. L'un des principaux problèmes qui se posent lorsqu'un réseau comporte un très grand nombre d'appareils est la saturation des ressources qui sont forcément limitées. Il serait très intéressant d'analyser la possibilité de développer un

modèle de simulation en ns-3 qui puisse prendre en charge simultanément des applications MTC massives et URLLC. Cela pourrait être réalisé en découpant les ressources du réseau et en les allouant par différentes tranches aux applications en fonction de leurs besoins. Un tel modèle de simulation développé dans ns-3 serait très bénéfique pour la communauté des chercheurs.

Outre les orientations futures, l'auteur estime que les fonctionnalités développées au cours de cette thèse seraient également très intéressantes pour la communauté des chercheurs travaillant dans le même domaine. Par conséquent les techniques de réduction de la latence, la planification hybride, les TTI flexibles et les modules de NB-IoT seront rendus publics. Le code sera également soumis à l'examen des modérateurs du projet ns-3 pour le module LTE en vue d'une éventuelle intégration dans la version principale du module LTE avec l'espoir que cela profitera aux développements scientifiques dans le domaine de l'URLLC et des réseaux cellulaires.

6.2 Conclusion

One of the newer requirements for cellular networks is to enable the support for mission-critical IoT applications. The increasing demand from such applications with regard to latency poses a challenge for cellular networks. To meet low latency requirements, there have been multiple approaches proposed and discussed in the recent past. In this thesis, the work towards latency reduction in 4G and 5G cellular networks has been performed by proposing, developing and evaluating multiple latency reduction techniques for user-plane latency. More precisely, the research questions outlined in section 1.7 have been answered in this thesis in the following way.

- **Q1. What are the inherent limitations of LTE with respect to achievable latency?** The radio access network latency in 4G LTE networks consists of two parts, control-plane and user-plane latency. The control plane latency mainly comes from the random access procedure. This procedure is mandatory for a User Equipment (UE) to register to the network and to gain the right to be included by the scheduling operated by the eNB. The user-plane latency depends on scheduling type, transmission time interval and processing time at the node. This latency comes for the necessity of the eNB to play its role of scheduler and to allow each connected UE a sending slot. In Chapter 3 of this thesis, a comprehensive analysis of both type of latencies is given. The theoretical analysis is supported by the evaluations using open-source simulator ns-3. This evaluation allows to calibrate and validate the code of the simulator for upcoming more detailed simulations. The analyses of LTE latency reveal that the minimum best case achievable user-plane latency is between 11 ms to 13 ms. The URLLC use case latency requirements given in Chapter 1 require latency of less than 10 ms for some application. Therefore, this analysis indicates the need for latency reduction in 4G LTE standards.

- **Q2. Which approaches are best suited to optimize latency of cellular networks in practice?** The transition from 4G LTE to 5G networks does not include a complete swap of infrastructure and services. It is envisioned that 4G LTE will also be part of the next generation 5G networks where many of use case requirements can still be fulfilled by the LTE networks. In order to support URLLC applications with 4G cellular networks, some of the potential latency reduction techniques have been investigated in this thesis. These techniques include short transmission time interval and semi-persistent scheduling. A major hurdle in evaluating these techniques was the lack of availability of these features in any of open-source simulators with LTE module. Therefore, these

features were developed for ns-3 simulator as part of work in this thesis and evaluated through simulations to point out the potential improvements they offer. The evaluation from the simulations as presented in Chapter 4 show very promising results from these techniques. It is noted that a combination of sTTI and SPS can reduce the uplink user-plane latency by 80%. However, both techniques also bring drawbacks of additional control overhead and resource wastage into the overall efficiency of the system.

- **Q3. How can the minimally achievable latency be achieved as functions of system characteristics?** The latency reduction techniques evaluated in Chapter 4 show very promising results for low-latency applications. It is also essential to evaluate these techniques for realistic use case scenarios to find out the limitations and amount of improvement they offer. To this end, both latency reduction techniques developed in ns-3 were evaluated over realistic simulation settings for the use cases mentioned in Chapter 1 of this thesis. The evaluation of use case scenarios indicated that both techniques could be used in 4G LTE systems with a limitation on the application data rate and device density. These limitations also point out the fact that the latency reduction techniques need further enhancements to support higher device densities and a diverse set of URLLC applications.

- **Q4. Which beyond-the-standards approaches are potential candidates for improving network performance, and in particular latency?** The evaluation of latency reduction techniques also shows the drawbacks of these techniques. In order to eliminate the effect of these drawbacks, two novel approaches are proposed and implemented in ns-3 simulator and evaluated through extensive simulations. The results show that presented approaches can overcome the drawbacks of latency reduction techniques. It is expected that these approaches can pave the way for 4G cellular networks in allowing support for URLLC applications.

Along with the improvements for latency, the latency reduction techniques (sTTI and SPS) also have their disadvantages as discussed in section 4.6. One of the major advantage of conventional dynamic scheduling in LTE is the efficient utilization of network resources. However, SPS does not keep the resource utilization efficiency and at a certain point the resources are wasted. The sTTI also decreases the system efficiency in terms of control overhead. Since the control signals in LTE are designed for legacy 14-os TTI, their transmission with sTTI brings extra control overhead, which eventually degrades system performance, such as reduced overall data rate. To overcome these drawbacks, in Chapter 5 two novel approaches have been proposed and evaluated. The Flexible TTI is designed to reduce the impact of sTTI

on system performance by using multi-sized (14-os, 7-os, and 2-os) TTI at the same time. On the other hand, hybrid scheduling aims to reduce the drawbacks of SPS by simultaneously utilizing the dynamic and the semi-persistent scheduling.

6.3 Outlook

There are a few directions, which could be investigated as future work for this thesis as following.

- Integration of RAN latency reduction techniques (from this thesis) with the next generation core networks to analyze and investigate the overall end-to-end latency of cellular networks. Since core network and Internet delays vary, it would be very interesting to include the overall system blocks in the latency improvement studies.
- One of the applications of 5G is massive MTC where the network is expected to support very large device densities. One of the major problem that arrives with having large number of devices in the network is the saturation of limited resources. It would be very interesting to analyze the possibility of developing a simulation model in ns-3, which can support both massive MTC and URLLC application simultaneously. This could be achieved by slicing the network resources and allocating the slices to applications according to their requirements. Such a simulation model developed in ns-3 would be highly beneficial for the research community.

Apart from the future directions, the features developed during this thesis would also be very interesting for the research community working in the similar field. Therefore, the latency reduction techniques, hybrid scheduling, flexible TTI, and NB-IoT features/modules will be made public for others through the public repository[1] once the thesis is published. The code will also be submitted to ns-3 LTE module moderators for the review for possible inclusion in the main LTE module release with a hope that this will greatly benefit the scientific developments in the field of URLLC and cellular networks.

As discussed in Chapter 2, there are other wireless technologies such as Wi-Fi or LPWAN that might also be interesting for some of the use cases presented in Chapter 1, Table 1.2. Similar activities to the ones presented in this thesis, could also be carried out for those wireless technologies, especially to support mMTC applications.

[1] https://github.com/hso-esk/

Bibliography

[1] "TS 22.261 Service requirements for the 5G system," 3GPP, 2017.

[2] "TR 25.913 Requirements for Evolved UTRA (E-UTRA) and Evolved UTRAN (E-UTRAN)," 3GPP, 2010.

[3] "Quality of Service Regulation Manual", Telecommunication development sector ITU, 2017.

[4] R. Ratasuk et al., “Performance of Low-Cost LTE Devices for Advanced Metering Infrastructure,” Proc. IEEE VTC, June 2013, pp. 1–5.

[5] Frotzscher, Andreas, et al. "Requirements and current solutions of wireless communication in industrial automation." Communications workshops (ICC), 2014 IEEE international conference on. IEEE, 2014.

[6] Parvez, Imtiaz, et al. "A survey on low latency towards 5G: RAN, core network and caching solutions." IEEE Communications Surveys & Tutorials (2018).

[7] Latif, Siddique, et al. "How 5g wireless (and concomitant technologies) will revolutionize healthcare?." Future Internet 9.4 (2017): 93.

[8] "3rd Generation Partnership Project", Retrieved January 15, 2020, from https://www.3gpp.org/

[9] K. Takeda et al. "Industry Perspectives: Latency Reduction Towards 5G" IEEE Wireless Communications 24.3 (2017): 2-4.

[10] 3GPP TR 45.820, ”Cellular system support for ultra-low complexity and low throughput Internet of Things (CIoT)”.

[11] Nashwan, Shadi. "SAK-AKA: A Secure Anonymity Key of Authentication and Key Agreement protocol for LTE network." International Arab Journal of Information Technology (IAJIT) 14.5 (2017).

[12] Taleb, Tarik, and Andreas Kunz. "Machine type communications in 3GPP networks: potential, challenges, and solutions." IEEE Communications Magazine 50.3 (2012): 178-184.

[13] “Worldwide Cellular M2M Modules Forecast Market Brief,” Beecham Research, Aug. 2010.

[14] S. Lucero, “Maximizing Mobile Operator Opportunities in M2M: The Benefits of an M2M-Optimized Network,” ABI Research, 1Q 2010.

[15] M. Frodigh. "Machine-type communication in 5G.", Retrieved February 17 2020, from https://www.ericsson.com/en/blog/2014/2/machine-type-communication-in-5g

[16] ITU-R, “Minimum requirements related to technical performance for IMT-2020 radio interface(s),” ITU - International Telecommunication Union, Tech. Rep., Nov. 2017.

[17] EventHelix, "Ultra-Reliable Low-Latency Communication (URLLC)", Retrieved September 23, 2019, from https://medium.com/5g-nr/ultra-reliable-low-latency-communication-urllc-9b2505e81579

[18] Dahlman, Erik, et al. "5G wireless access: requirements and realization." IEEE Communications Magazine 52.12 (2014): 42-47.

[19] Z. Li, M. A. Uusitallo, H. Shariatmadari, B. Singh, “5G URLLC: Design Challenges and System Concepts,” 15th International Symposium on Wireless Communication Systems, pp. 1–6, 2018.

[20] Janne Peisa, "5G Techniques for Ultra Reliable Low Latency Communication", Retrieved September 23, 2019, from http://cscn2017.ieee-cscn.org/files/2017/08/Janne_Peisa_Ericsson_CSCN2017.pdf

[21] Chen, He, et al. "Ultra-reliable low latency cellular networks: Use cases, challenges and approaches." IEEE Communications Magazine 56.12 (2018): 119-125.

[22] N. Larson et al., “Investigating Excessive Delays in Mobile Broadband Networks,” Proc. SIGCOMM Workshops Things Cellular, Aug. 2015, pp. 51–56.

[23] Nsiah, Kofi Atta, et al. "Performance Evaluation of Latency for NB-LTE Networks in Industrial Automation." 2019 IEEE 30th Annual International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC). IEEE, 2019.

[24] Zhang, Junping, et al. "Data-driven intelligent transportation systems: A survey." IEEE Transactions on Intelligent Transportation Systems 12.4 (2011): 1624-1639.

[25] Haider, Amir, and Seung-Hoon Hwang. "Adaptive Transmit Power Control Algorithm for Sensing-Based Semi-Persistent Scheduling in C-V2X Mode 4 Communication." Electronics 8.8 (2019): 846.

[26] 5G PPP, "5G Automotive Vision" white paper, 2015.

[27] "Self-Driving.. You Out of Business", Retrieved 19th February, 2020 from https://digital.hbs.edu/platform-rctom/submission/self-driving-you-out-of-business/

[28] Ashraf, Shehzad A., et al. "Ultra-reliable and low-latency communication for wireless factory automation: From LTE to 5G." Emerging Technologies and Factory Automation (ETFA), 2016 IEEE 21st International Conference on. IEEE, 2016.

[29] "IRIMAS – INSTITUT DE RECHERCHE EN INFORMATIQUE, MATHÉMATIQUES, AUTOMATIQUE ET SIGNAL", Retrieved March 25th, 2020, from https://www.irimas.uha.fr/

[30] "Institut für verlässliche Embedded Systems und Kommunikationselektronik", Retrieved March 25th, 2020, from https://ivesk.hs-offenburg.de/

[31] 3GPP TR 38.913 v15.0.0, “Study on scenarios and requirements for next generation access technologies”, Sept 2018.

[32] Lema, M.A., Laya, A., Mahmoodi, T., Cuevas, M., Sachs, J., Markendahl, J. and Dohler, M., 2017. "Business Case and Technology Analysis for 5G Low Latency Applications." IEEE Access, 5, pp.5917-5935.

[33] “ns-3.” Retrieved September 23, 2019, from http://www.nsnam.org

[34] A. Virdis, G. Stea, G. Nardini, "Simulating LTE/LTE-Advanced Networks with SimuLTE", DOI 10.1007/978-3-319-26470-7_5, in: M.S. Obaidat, J. Kacprzyk, T. Ören, J. Filipe, (eds.) "Simulation and Modeling Methodologies, Technologies and Applications", Volume 402 of the series Advances in Intelligent Systems and Computing, pp. 83-105, Springer, ISBN 978-3-319-26469-1, 15 January 2016

[35] "OMNet++", Retrieved September 23, 2019, from https://omnetpp.org

[36] "LTE toolbox", Retrieved September 23, 2019, from https://www.mathworks.com/products/lte.html

[37] "Massive Machine Type Communication in 5G and beyond network", Retrieved February 18, 2020, from https://itnspotlight.com/massive-machine-type-communication-in-5g-and-beyond-network/

[38] Arun Kumar, "Introduction to LORA technology", Retrieved September 23, 2019, from http://www.embien.com/blog/introduction-to-lora-technology/

[39] É. Morin M. Maman R. Guizzetti and A. Duda. Comparison of the device lifetime in wireless networks for the internet of things. IEEE Access, 2017 (cit. on p. 6).

[40] B. Martinez F. Adelantado X. Vilajosana P. Tuset-Peiro. Understanding the limits of lorawan. IEEE Communications Magazine, 2017 (cit. on p. 6).

[41] Mekki, Kais, et al. "A comparative study of LPWAN technologies for large-scale IoT deployment." ICT express 5.1 (2019): 1-7.

[42] J. J. Nielsen R. B. Sørensen D. M. Kim and P. Popovski. Analysis of latencyand mac-layer performance for class a lorawan. IEEE Wireless Communica- tions Letters, 2017 (cit. on p. 8).

[43] J. Zou M. Hua T. Xia W. Yang M. Wang J. Zhang and X. You. Narrowband wireless access for low-power massive internet of things: A bandwidth perspective. IEEE Wireless Communications, 2017 (cit. on pp. 7, 8).

[44] J. de Carvalho Silva J. Rodrigues A. M. Alberti P. Solic and A. L. Aquino. Lorawan a low power wan protocol for internet of things: A review and opportunities. in Computer and Energy Science (SpliTech) 2nd International Multidisciplinary Conference on.IEEE, 2017 (cit. on pp. 6, 7).

[45] Sigfox. Sigfox technology overview. Retrieved September 23, 2019, from https://www.sigfox.com/en/sigfox-iot-technology-overview (cit. on p. 8).

[46] J. Sebastian E, A. Sikora, M. Schappacher, and Z. Amjad. "Test and Measurement of LPWAN and Cellular IoT Networks in a Unified Testbed." 2019 IEEE 17th International Conference on Industrial Informatics (INDIN). Vol. 1. IEEE, 2019.

[47] Fraunhofer IIS, "MIOTY. The wireless IoT technology", Retrieved December 6th, 2019, from https://www.iis.fraunhofer.de/en/ff/lv/net/tech/telemetrie.html

[48] WiFi (Wireless Fidelity), Retrieved March 23rd, 2020, from http://www.wi-fi.org

[49] "Industrial Wireless, selecting a wireless technology", Retrieved March 23rd, 2020, from http://www.bb-elec.com/Learning-Center/All-White-Papers/Wireless-Cellular/Industrial-Wireless-Selecting-a-Wireless-Technolog.aspx

[50] G.Santandrea, “A PROFINET IO application implemented on Wireless LAN,” in IEEE International Workshop on Factory Communication Systems, Turin, Italia, Jun. 2006.

[51] "IWLAN – the WLAN for challenging industrial applications", Retrieved April 12th, 2020, from https://new.siemens.com/global/en/products/automation/industrial-communication/industrial-wireless-lan.html

[52] Gidlund, Mikael, Tomas Lennvall, and Johan Åkerberg. "Will 5G become yet another wireless technology for industrial automation?." 2017 IEEE International Conference on Industrial Technology (ICIT). IEEE, 2017.

[53] IEEE Standards Association. "IEEE standard for local and metropolitan area networks-specific requirements: Part 11 Wireless LAN MAC and PHY specifications Amendment 6: Wireless Access in Vehicular Environments", available at: http://standards. ieee. org/.

[54] Hernandez-Jayo, Unai, and Idoia De-la-Iglesia. "Reliability of Cooperative Vehicular Applications on Real Scenarios Over an IEEE 802.11 p Communications Architecture." International Conference on E-Business and Telecommunications. Springer, Berlin, Heidelberg, 2013.

[55] Patel, Maulik, and Vijay Ukani. "Optimized handoff process in IEEE802. 11p based VANET." 2012 1st International Conference on Emerging Technology Trends in Electronics, Communication & Networking. IEEE, 2012.

[56] Amjad, Zubair, et al. "Low Latency V2X Applications and Network Requirements: Performance Evaluation." 2018 IEEE Intelligent Vehicles Symposium (IV). IEEE, 2018.

[57] 3GPP TR 36.888, "Study on provision of low-cost Machine-Type Communications (MTC) User Equipments (UEs) based on LTE", v.12.0.0, June 2013.

[58] 3GPP RP-150492, "Revised WI: Further LTE Physical Layer Enhancements for MTC," Ericsson, RAN67, Shanghai, China.

[59] 3GPP TS 36.306 v14.3.0, "Evolved Universal Terrestrial Radio Access (E-UTRA); User Equipment (UE) radio access capabilities (Release 14)," June. 2016.

[60] Ghavimi, Fayezeh, and Hsiao-Hwa Chen. "M2M communications in 3GPP LTE/LTE-A networks: Architectures, service requirements, challenges, and applications." IEEE Communications Surveys & Tutorials 17.2 (2014): 525-549.

[61] R. Ratasuk N. Mangalvedhe A. Ghosh and B. Vejlgaard. Narrowband lte-m system for m2m communication. in Vehicular Technology Conference (VTC Fall), 2014 (cit. on p. 9).

[62] Raquel Ligero, "Differences between Nb-IoT and LTE-M", Retrieved September 23, 2019, from https://accent-systems.com/blog/differences-nb-iot-lte-m/

[63] Bi, Qi. "Ten Trends in the Cellular Industry and an Outlook on 6G." IEEE Communications Magazine 57.12 (2019): 31-36.

[64] P. Schulz, et al. “Latency Critical IoT Applications in 5G: Perspective on the Design of Radio Interface and Network Architecture ”. IEEE Communication Magazine , Volume 55, Issue 2, pp. 70–78, Feb. 2018

[65] Aktas, Ismet, et al. "LTE evolution—Latency reduction and reliability enhancements for wireless industrial automation." Personal, Indoor, and Mobile Radio Communications (PIMRC), 2017 IEEE 28th Annual International Symposium on. IEEE, 2017.

[66] 3GPP TR 36.912 v14.0.0, "Feasibility study for Further Advancements for E-UTRA (LTE-Advanced)".

[67] "5G technology in industrial campus networks", Retrieved March 25th, 2020, from https://www.telekom.com/en/company/details/5g-technology-in-campus-networks-556692

[68] N. Baldo, M. Miozzo, M. Requena, J. Nin Guerrero, An Open Source Product-Oriented LTE Network Simulator based on ns-3, in Proceedings of the 14th ACM International Conference on Modeling, Analysis and Simulation of Wireless and Mobile Systems (MSWIM 2011), 31-4 November 2011, Miami Beach, FL (USA).

[69] "NS-3 LTE design documentation", Retrieved September 23, 2019, from https://www.nsnam.org/docs/models/html/lte-design.html

[70] “LTE-EPC Network simulAtor.” Retrieved September 23, 2019, from http://networks.cttc.es/mobile-networks/software-tools/lena/

[71] Dawaliby, Samir, Abbas Bradai, and Yannis Pousset. "In depth performance evaluation of LTE-M for M2M communications." Wireless and Mobile Computing, Networking and Communications (WiMob), 2016 IEEE 12th International Conference on. IEEE, 2016.

[72] Sesia, Stefania, Matthew Baker, and Issam Toufik. LTE-the UMTS long term evolution: from theory to practice. John Wiley & Sons, 2011.

[73] Polese, M., Centenaro, M., Zanella, A. and Zorzi, M., 2016, May. ”M2M massive access in LTE: RACH performance evaluation in a Smart City scenario.” In Communications (ICC), 2016 IEEE International Conference on (pp. 1-6). IEEE.

[74] Laya, A., Alonso, L. and Alonso-Zarate, J., 2014. "Is the Random Access Channel of LTE and LTE-A Suitable for M2M Communications? A Survey of Alternatives." IEEE Communications Surveys and Tutorials, 16(1), pp.4-16.

[75] "NB-IoT NS-3 development", Retrieved September 23, 2019, from https://www.nsnam.org/wiki/NB-IOT

[76] Ashish Kumar Sultania, Pouria Zand, Chris Blondia, Jeroen Famaey. "Energy Modeling and Evaluation of NB-IoT with PSM and DRX", in IEEE Global Communications Conference: Workshops: Green and Sustainable 5G Wireless Networks, 9-13 Dec 2018.

[77] 3GPP TR 36.881 v0.6.0, "Evolved Universal Terrestrial Radio Access (E-UTRA); Study on Latency Reduction Techniques for LTE".

[78] 3GPP RP-160667, "Work Item on L2 Latency Reduction Techniques for LTE," March 2016.

[79] 3GPP RP-161299, "Work Item on Shortened TTI and Processing Time for LTE," June 2016.

[80] Amjad, Zubair, et al. "Latency Reduction for Narrowband LTE with Semi-Persistent Scheduling." 2018 IEEE 4th International Symposium on Wireless Systems within the International Conferences on Intelligent Data Acquisition and Advanced Computing Systems (IDAACS-SWS). IEEE, 2018.

[81] Amjad, Zubair, et al. "Latency Reduction in Narrowband 4G LTE Networks." 2018 15th International Symposium on Wireless Communication Systems (ISWCS). IEEE, 2018.

[82] Kofi Atta Nsiah, et al., "Performance Evaluation of Ultra-Low Latency Wireless Communication in Industrial Automation", Embedded World 2018.

[83] Kofi Atta, et al. "Latency Reduction Techniques for NB-IoT Networks", IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems (IDAACS), IEEE, 2019.

[84] Luoto, P., Bennis, M., Pirinen, P., Samarakoon, S., Horneman, K. and Latva-aho, M., 2017, June. "Vehicle clustering for improving enhanced LTE-V2X network performance." In Networks and Communications (EuCNC), 2017 European Conference on (pp. 1-5). IEEE.

[85] Sikora, Axel, and Manuel Schappacher. "A highly scalable IEEE802. 11p communication and localization subsystem for autonomous urban driving." Connected Vehicles and Expo (ICCVE), 2013 International Conference on. IEEE, 2013.

[86] Ledy, J., Poussard, A. M., Vauzelle, R., Hilt, B., and Boeglen, H. (2012, August). "AODV enhancements in a realistic VANET context." In Wireless Communications in Unusual and Confined Areas (ICWCUCA), 2012 International Conference on (pp. 1-5). IEEE.

[87] Chen, S., Hu, J., Shi, Y., Peng, Y., Fang, J., Zhao, R. and Zhao, L., 2017. "Vehicle-to-Everything (v2x) Services Supported by LTE-Based Systems and 5G." IEEE Communications Standards Magazine, 1(2), pp.70-76.

[88] Thomesse, J-P. "Fieldbus technology in industrial automation." Proceedings of the IEEE 93.6 (2005): 1073-1101.

[89] Hoymann, Christian, et al. "LTE release 14 outlook." IEEE Communications Magazine 54.6 (2016): 44-49.

[90] J. C. S. Arenas, T. Dudda, L. Falconetti, "Ultra-low latency in Next Generation LTE Radio Access", 11th International ITG Conference on Systems, Communications and Coding, pp. 1–6. , 2017.

[91] G. Pocovi et al. "On the impact of multi-user traffic dynamics on low latency communications" Wireless Communication Systems (ISWCS), 2016 International Symposium on. IEEE, 2016.

[92] J. Li, H. Sahlin, G. Wikström, "Uplink PHY Design with Shortened TTI for Latency Reduction", IEEE Wireless Communications and Networking Conference, pp.1–5, 2017.

[93] Hosseini, Kianoush, et al. "Link-Level Analysis of Low Latency Operation in LTE Networks." Global Communications Conference (GLOBECOM), 2016 IEEE. IEEE, 2016.

[94] K. Hosseini, S. Patel, A. Damnjanovic, W. Chen, J. Montojo, "Link-Level Analysis of Low Latency Operation in LTE networks", IEEE Global Communications Conference , pp. 1–6 , 2016.

[95] S. Xiaotong, H. Nan, Z. Naizheng, "Study on system latency reduction based on shorten TTI", IEEE 13th International Conference on Signal Processing (ICSP), pp. 1293–1297. , 2016

[96] X. Zhang, "Latency reduction with short processing time and short TTI length", International Symposium on Intelligent Signal Processing and Communication Systems, pp. 1293–1297. , 2017

[97] Z. Zhang, Y. Gao, Y. Liu, Z. Li, "Performance Evaluation of Shortenend Transmission Time Interval in LTE networks", IEEE Wireless Communications and Networking Conference, pp. 1–5, 2018

[98] Lee, Kwonjong, et al. "Latency of cellular-based V2X: Perspectives on TTI-proportional latency and TTI-independent latency." IEEE Access 5 (2017): 15800-15809.

[99] Fehrenbach, Thomas, et al. "URLLC Services in 5G Low Latency Enhancements for LTE." 2018 IEEE 88th Vehicular Technology Conference (VTC-Fall). IEEE, 2018.

[100] Pocovi, Guillermo, et al. "Achieving ultra-reliable low-latency communications: Challenges and envisioned system enhancements." IEEE Network 32.2 (2018): 8-15.

[101] M. Li, X. Guan, C. Hua, C. Chen, L. Lyu, “Predictive Pre-allocation for Low-Latency Uplink Access in Industrial Wireless Networks”, IEEE Conference on Computer Communications, 2018.

[102] N. Afrin, J. Brown, J. Y. Khan, “Performance evaluation of an adaptive semi-persistent LTE packet scheduler for M2M communications”, 8th International Conference on Signal Processing and Communication Systems (ICSPCS), 2018.

[103] Ali, Samad, Nandana Rajatheva, and Walid Saad. "Fast uplink grant for machine type communications: Challenges and opportunities." IEEE Communications Magazine 57.3 (2019): 97-103.

[104] Abreu, Renato, Preben Mogensen, and Klaus I. Pedersen. "Pre-scheduled resources for retransmissions in ultra-reliable and low latency communications." 2017 IEEE Wireless Communications and Networking Conference (WCNC). IEEE, 2017.

[105] Pocovi, Guillermo, et al. "MAC layer enhancements for ultra-reliable low-latency communications in cellular networks." 2017 IEEE International Conference on Communications Workshops (ICC Workshops). IEEE, 2017.

[106] "NS-3 LTE user documentation", Retrieved September 23, 2019, from https://www.nsnam.org/docs/models/html/lte-user.html#configure-lte-mac-scheduler

[107] Rayal. F., "LTE Peak Capacity Explained: How to Calculate it?", Retrieved December 3rd, 2019, from https://frankrayal.com/2011/06/27/lte-peak-capacity

[108] "Simulation of Urban Mobility (SUMO)", Retrieved April 8th, from https://sumo.dlr.de