REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE
MINISTERE DE L'ENSEIGNEMENT SUPERIEUR ET DE LA RECHERCHE SCIENTIFIQUE
Université Amar Telidji - LAGHOUAT –

Faculté des Sciences et technologies

Département Electronique

Mémoire de Master en Télécommunications

Option : Systèmes De Télécommunications

THEME

Reducing Latency In 5G networks

Presented by :

- *Recioui mouadh*
- *Elabed sofiane*

jury of Soutenance :

Name	**Grade**	**quality**
Challali Safouane	MCB	Advisor
RAMDANI Saadi	MAA	president of jury
MERAH hocine	MCB	a jury member

Abstract

One of the most challenges is to fulfil the whole essential requirement for a 5G system, and to understand the tricks for obtaining a lower latency which is the most dominant aspect in 5G, since the world had enough bandwidth in the previous generations for daily uses. However, latency was not a big concern, but for today applications such as VANET and Online gaming, the fact that control or the location update information depends more on the latency rather than the throughput, therefore, in this desertion, we would target this aspect, at first we would understand the evolution to 5G NR based on the needs of humans and how technology fought to accomplish its mission, understand KPI for evaluating a cellular system. We would go more in-depth into much more technical aspects; at first, we divide the latency in each part of the network. We would focus on the RAN latency, especially how waveform and frame flexibility structure, channels and duplex modes effects latency with numerologies

Keywords : 5G, latency, frame structure, 5G NR, KPI.

واحدة من اهم المفاتيح لتجسيد نظام الجيل الخامس في الواقع هو وقت الاستجابة, احتياجات الانسان العادي ورضاه عن سرعة الانترنت شبه كاف لمختلف التطبيقات المعهودة كمشاهدة فيلم او الاتصال عبر الانترنت او تحميل من الانترنت والتي قامت بتوفيرها خدمات الجيل الرابع , لكن على رغم من ذلك فان وقت الاستجابة الذي لم يكن مهما في الاجيال السابقة بقدر ماهو الان في الجيل الخامس اصبح الاهم , خاصة في مجال الالعاب و السيارات ذاتية القيادة لأنه حتى وان كانت كمية بيانات التحكم او تحديث موقع عبر احداثيات تعتبر بايانات جد صغيرة الا انها تحتاج تحديث بسرعة كبيرة , وهذا ما سوف نحاول شرحه وفهمه في هذه الاطروحة .بدا من تطور شبكات الهاتف النقال حتى وصولها الى الجيل الخامس, وفهم ماهية العراقيل التي لعبت دور كبير في تطور الشبكة والوصول الى هدف المنشود وهو تحقيق الرضى الكافي للمستخدم الاخير .بعد ذلك سوف ندخل في مفاهيم تقنية اكثر و نتحدث عن وقت الاستجابة في شبكة الهاتف ونقسم هذه الاوقات على كل الوحدات الموجودة .وفي دراستنا هذه سوف نركز اكثر على وقت الاستجابة من المستعمل الى برج الاتصال وخاصة دراسة اطار البيانات وشرح دوره الكبير في تقليل وقت الاستجابة بالنسبة لنظام الجيل الخامس

الكلمات المفتاحية : الجيل الخامس , وقت الاستجابة ,شبكات الهاتف النقال .

L'une des clés les plus importantes pour incarner la 5eme génération (5G) en réalité est le temps de réponse, ou la latence , les besoins des utilisateurs et la satisfaction à l'égard de la vitesse Internet presque suffisante pour les diverses applications habituelles telles que regarder un film ou communiquer via Internet ou télécharger à partir d'Internet qui a fourni les services de quatrième génération (4G) , mais malgré la latence n'était pas vraiment importante dans les générations précédentes (25ms est suffisante) autant que dans la cinquième génération, la latence est devenue la plus importante, en particulier dans le domaine des jeux online et des voitures autonomes, car même si la quantité de données de contrôle ou de mise à jour des des coordonnées GPS est considérée comme de très petites données, elle doit être mise à jour très rapidement, et c'est ce que Nous essaierons de l'expliquer et de le comprendre dans cette thèse. En démarre par développement des réseaux mobiles jusqu'à la cinquième génération, et de comprendre quels obstacles ont joué un rôle majeur dans le développement du réseau et d'atteindre l'objectif souhaité qui est d'obtenir une satisfaction adéquate pour le dernier utilisateur, en comprend l'évolution vers La 5G basée sur les besoins des humains et la façon dont la technologie s'est battue pour accomplir sa mission, comprendre les KPI pour évaluer un système cellulaire, puis on vas etuider la trame en têrme physique , on fait une simulation a la latence des packets utilisant plusieurs espacement des porteuses .

Mot clés : 5G, la latence ,kpi, la trame .

Acknowledgement

This work would not have been possible without, **ALLAH** in the first place, family members, especially my dear **mother** and **father**, in addition to **Dr CHELLALI SAFOUANE** who have been supportive of my dessertation goals and who worked actively to provide me with the right notes about my work. I am grateful to all of those with whom I have had the pleasure to work during my five years of study, thanks to the telecommunication community in social media, which was one of the most source of motivation to study and thrive for knowledge.

Secondly, I extend my gratitude to every teacher who taught us, especially **administration staff,** for every single effort they did, I believe that without them, things will get lot harder.

I am also hugely appreciative to my right hand, my partner in this dissertation, who worked so hard in gathering all the sources of information about this new topic, as well as giving the ideas and the critics about which really helped to improve the quality of the work.

Finally, but by no means least, thanks go to OmAR AiT AISsA, MuStApha BeNgAnA , ElaBeD SoUfiaNe and all the friends to their unbelievable support.

‘‘ *وَقَوْلِي كلَّما جَشَأَتْ وجاشَتْ مَكانَكِ، تُحْمَدِي أَو تَسْتَرِيحِي* ” عمرو بن الإطنابة.

‘‘*he who never fails, never tried something new*’’ albert Einstein.

Mes remerciements. ありがとう

Table Of contents

List of Figures

List of Tables

(AD/DA)C: analog-to-digital/digital-to-analog converter
(D)STBC: (Distributed) Space Time Block Coding
(L)RAP: (Light) Radio access point
(N)LOS: (NON) Line Of sight
(N)SA: (NON)-Standalone
(S/E)E: spectrum/energy efficiency
(e)MBMS: (evolved)Multimedia Broadcast Multicast Services
(m) MIMO: (massive) Multiple-Input Multiple-Output
3GPP: 3rd Generation Partnership Project
5GC: 5G core
API: application programming interfaces
AWGN: additive white gaussian noise
BBU: baseband unit
BER: bit error rate
BS-BS: baste station to base station
BS: base station
BWP: Bandwidth Part
C-RAN: cloud-radio access technology
CCE: control channel elements
CDMA: Code-Division Multiple Access
CN: core network
CP: cyclic prefix
CRC: cyclic redundancy check
CS: circuit switching
CSI-RS: Channel State Information Reference Signal
CSI: channel state information
CoMP: Coordinated multipoint
D2D: Device to Device
DCI: downlink channel indicator
DFTS-OFDMA: DFT spread OFDM
DFTS: discreet Fourier transform
DL: downlink
DMRS: demodulation reference signal
DUEs: D2D UEs
E-UTRAN: evolved UMTS terrestrial radio access network
E2E : end to end
EED: end to end delay
EPC: evolved packet core
EU FP7 : Europe seventh framework programming
EVM : error vector magnitude
FBMC: filter bank multi-carrier
FDD: frequency Division Duplex
FSPL: free space path loss
GPRS: general packet radio service
GSM: Global System for Mobile
GTP-U: tunneling protocol for user plane
HARQ: Hybrid automatic repeat request
HSPA+ : High Speed Packet Access plus
ICIC: Inter-cell interference coordination
ICT: Information and communication technology
IEEE: Institute of Electrical and Electronics Engineers

IMT-Advanced: International Mobile Telecommunications-Advanced
IoT: Internet of Things
KPI: Key Performance Indicator
LCS: location services
LDPC: low-density parity-check
LLC: Low Latency Communication
LTE-A: Long Term Evolution- Advanced
LTE: Long Term Evolution
MBB: Mobile Broadband
MCPTT: Mission Critical Push-To-Talk
MEC: mobile edge computing
MME: Mobile Management Entity
MS: mobile station
MSC : mobile switching center
MTC: Machine Type Communication
MU-MIMO: multi-user MIMO
NFV: network function virtualization
NOMA: Non-orthogonal multiple access
NR: new radio
OFDMA: orthogonal frequency division multiple access
OOB: out of band
PA: power amplifier
PAPR: peak-to-average power ratio
PCRF: policy and charging rules function
PD(V/R): packet delay variation/ratio
PDCCH: Physical Downlink Control Channel
PDN: Packet Data Network
PDSCH: Physical Downlink Shared Channel
PRB: physical resource block
PS: packet switching
PTRS: Phase-Tracking Reference Signals
PUCCH: Physical Uplink Control Channel
PUSCH: Physical uplink Shared Channel
PWS: Public Warning System
QAM: quadrature amplitude modulation
QPSK: quadrature pulse shift keying
QoE: quality of experience
QoS: Quality of service
RAN: Radio Access Network
RANaaS: radio access network as a service
RAT: Radio access technology
REG: Resource Element Groups
RF: radio frequency
RMS: root mean square
RN: Relay Node
RNC: radio network control
RRC: radio resource control
RRM: Radio resource management
RTT: round-trip time
RoF: radio-over-fiber

S-GW: Serving Gateway
SC-FDMA: Single- Carrier Frequency-Division Multiple Access
SCMA: sparse code multiple access
SCS: subcarrier spacing
SDN: software defined networking
SDU: service data unit
SGW-C: service gateway control plane
SIC : self-interference cancellation
SISO: single input single output
SNR: signal to noise ration
SON: Self-Organizing Networks
SR: scheduling request
SRS: Sounding Reference Signals
SU- MIMO: single-user MIMO
TD-SCDMA: time division spread code division multiple access
TDD: Time Division Duplex
TDM: time division multiplexing
TDMA: time-Division Multiple Access
TM: Test model
TTI: transmission time interval
UCE: unified control entity
UDW: unified data gateway
UE: user end
UL: uplink
UMTS: Universal Mobile Telecommunications Service
URLLC: Ultra-reliable and low latency communications
V2V: vehicle to vehicle
VLC: visible light communication
WCDMA: Wideband Code-Division Multiple Access
WiMAX: Worldwide Interoperability for Microwave Access
eMBB: enhanced mobile broadband
eNB: Evolved Node B
mMTC: massive Machine Type Communication
mmwave: millimeter wave

General introduction

Latency is becoming one of the most significant aspect in telecommunications nowadays ,as the evolution of communication systems ,some applications need a very critical latency in other hand, we have learnt in our college education career that a telecom system should fulfil the three aspects which are: the availability, integrity and the confidentiality. Thus, the absolute importance of the power of telecom always brings up the quote “the power of a country is measured by two factors: the army, and the telecom infrastructure” said by a teacher B.RIDA from university of Amar telidji.

Since telecom infrastructure has many fields, from a daily basis use to the industrial applications, due to the telecom, our lives are made even more comfortable. We can access our web services from any place in the world, we can communicate with our families, friends, co-workers without any boundaries, thus, in an industrial section, we can monitor and acquire data from the machine in the fields and do the different operations on this data.

Mobile or cellular network is the most telecom application that plays a significant role in our lives, basically, glad to the mobility that it provides us, However, this did not end here, starting from analog phone or the 1G which was not so popular, the 2G provided us with the mobility and many voice services which were the critical requirement of the cellular network back then, such as texting, call on-hold, MMS … ,the human lives isn’t only based on talking all the time ,especially after the big evolution of the internet, the realizes of the 2G didn’t fulfil the requirements and that can be to the old access technology which was TDMA ,the throughput provided wasn’t sufficient for a user to watch videos or to browse with ease ,the 3G came up with a solution for giving the accessibility to the web ,which was the next goal for the cellular communication system ,glad to the CDMA technology and separation of packet switching from the voice ,3G system had a long journey since everyone could access internet with reasonable throughput ,many people were fine with the service provided by this generation, but the world of technology isn’t only for telecom ,the revolution in other technologies made 3G embarrassed ,such as HD videos, Online gaming , “the big war is telecom and informatics already settled ,we no longer depend on the circuit switching in the voice communication ,the packet switching is doing the job “ said by R.Saadi from university of Amar telidji. That made the 4G the first full IP-based system, a sufficient throughput and latency for the evolution of digitalization, by introduction the OFDM approach which played the hero role in 4G systems, but the war of technology did not end here, the world is growing fast and one year of human lives equals four years of technology, a new concept of other technologies was introduced and forced the 4G system to seem old, but rather than throughput problem which was the big concern in the previous generations, now the latency has climbed the stairs and forced itself to be the most critical aspect, which was the primary focus of the 5G, achieving a 1ms latency in our communication system rather than 25ms in 4G, with a reliable link, it is the hardest challenge that the telecom infrastructure has ever seen, the VANET, D2D, tactile internet, telemedicine, internet of things .. And all the promised fictional applications are becoming real in the 5G cellular communication system.

The concept of latency is divided in many, every optimization in each section is essential, the evolution of networking made it even easier by introducing the SDN and NFV, in the core, as well as the variable numerologies in the new mmwave spectrum. In this desertion, we are going to study the latency of 5G system and focus on the RAN, Especially in how the frame structure of 5G plays a big role in reaching an ultra-reliable low latency system using numerologies .

Chapter I: The Road to 5G

I-1 Introduction

Road to the fifth generation has seen many developments and standardizations ,since every release came up with a solution for a better service and for fulfilling the requirements , We are now living in a world where business meetings, university online courses and long-distance medical assistance are considered to be part of our daily routine. We have more access to information than ever before, and All this is achievable related to advances and innovations in the field of wireless connectivity.

The MBB (Mobile Broadband) service corresponds to applications and services that require an ever faster connection, to allow for example to watch 2 LTE-Advanced Pro videos in ultra-high definition or to use virtual or augmented reality applications, the LLC (Low Latency Communication) service groups all the applications requiring extremely high reactivity as well as reliable service data transmission, such as civil security for critical missions, the MTC (Machine Type Communication) service mainly includes users related to the Internet of Things. These services do not require very high data rates, but require more extensive coverage and lower energy consumption. In this chapter we will discuss how was the road map from the previous ip-based cellular network to the 5G and see the main changes.

I-2 Evolution of LTE to 5G

The development of LTE began during the period of HSPA+ (High Speed Packet Access plus) release of 3GPP in December 2004 [1]. The first release of LTE is called the release 8 of 3GPP which was introduced in December 2008. The release 9 of 3GPP finalized in December 2009 is the second release of LTE. After that, LTE-Advanced came into the market in June 2011 by the 3GPP as release 10. The release 10 and beyond by the 3GPP are called LTE-Advanced. The brief description of LTE evolutions up to LTE-Advanced are summarized below:

I-2-1 LTE Release 8

The Release 8 of 3GPP is referred to as the first LTE standard. The deployment layout of first LTE release is mainly macro/microcell based layout. It can provide high peak data rates than earlier HSPA+. The system capacity and the area coverage has been improved. Other major features that improved in LTE release 8 are low latency, reduced operating costs, multi-antenna support, flexible bandwidth operation.[1]

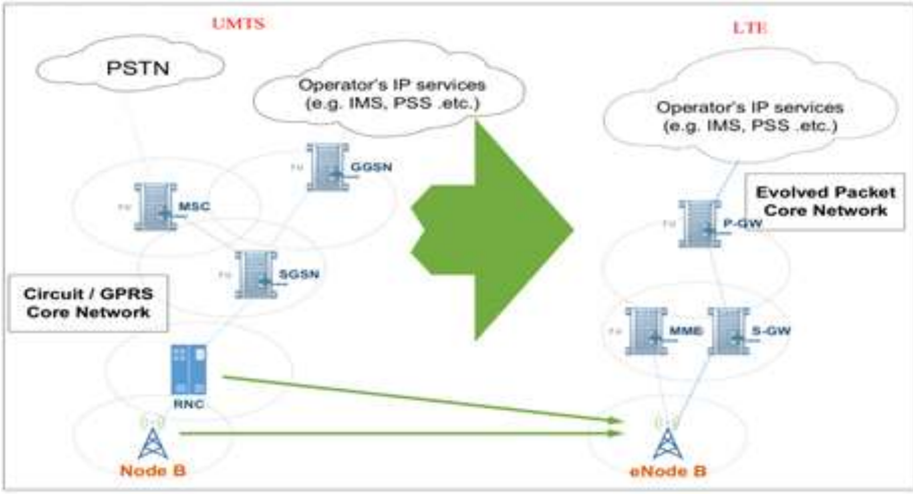


Figure I-1: GSM Access Network Vs LTE Access Network.

Figure I-1 above explains how LTE access network evolved from earlier GSM network architecture of UMTS. Node B and RNC are combined as e-node B in the LTE network. LTE uses OFDMA radio access technology and provides orthogonality between multiple users. Power control and inter-cell interference coordination is also an essential property of LTE.

I-2-2 LTE Release 9

LTE Release 9 is the complete release of LTE which was not completed in release 8. It includes broadcast/multicast services, positioning services, and enhanced emergency-call functionality.

- Multimedia Broadcast Multicast Services (MBMS) for LTE
- LTE MIMO: dual-layer beamforming
- LTE positioning
- PWS (Public Warning System)
- RF requirements for multi-carrier and multi-RAT base stations
- Home eNodeB specification (femto-cell)
- Self-Organizing Networks (SON).

Release 9 was finalized in the end of December 2009. The first commercial LTE was deployed in Sweden and Norway in 14 December, 2009. This deployment was the Release 9 of LTE, which provided interoperability between WiMAX (IEEE 802.16) and converged together the WCDMA of 3GPP and CDMA-2000 of 3GPP2. [1]

I-2-3 LTE Release 10

LTE Release 10, was finalized at the end of 2010 in which further improvements were added compared to release 8/9 in terms of performance and capabilities. LTE Release 10 and beyond are called LTE-Advanced which meet all the requirements of IMT-Advanced. So LTE-Advanced is a 4G mobile communication system which includes some additional features than previous releases. These features includes:

- Carrier aggregation
- Advanced MIMO techniques
- Wireless relaying
- Enhanced Inter-cell interference coordination (eICIC)
- Coordinated multipoint (CoMP) transmission/reception.

In carrier aggregation, multiple carrier components are aggregated to provide support for high transmission bandwidth. LTE-Advanced can support carrier aggregation of up to 100 MHz. Wireless relaying provides cell coverage extension and improved cell edge performance. [3].

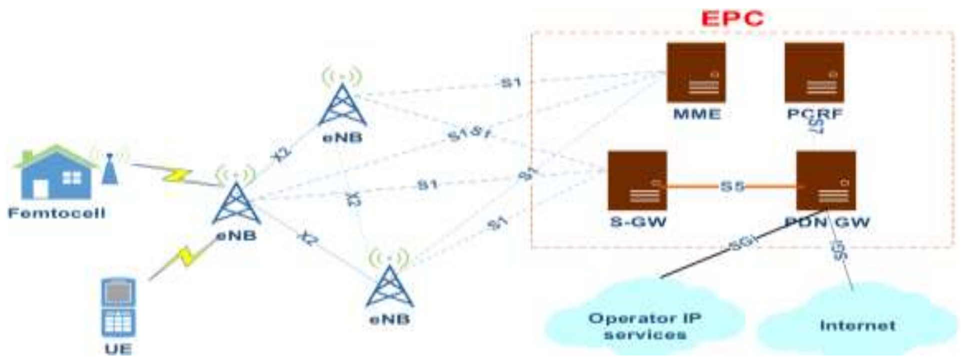


Figure I-2: LTE-Advanced Network Architecture.

The network architecture of LTE-Advanced is illustrated in figure 3. All of the eNBs (Evolved NodeB) are connected to the S-GW (Serving Gateway) and MME of EPC. eNB is the combination of UMTS (3G) NB and RNC (Radio Network Control) [2].

. A comparison study among LTE releases (Release 8 to 10) by considering what enhancements have been done in the newer versions is summarized in fig 1 below:

Table I-1: enhancements in LTE releases.

Enhancements in LTE releases		
Release-8	**Release-9**	**Release-10 (LTE-A)**
•Spectrum flexibility •Multi antenna transmission •ICIC	•Spectrum flexibility •Multi antenna transmission •ICIC •LTE positioning •broadcast/multicast t services (MBMS) •Home eNodeB specification (femto-cell) •LTE MIMO: dual layer beamforming,	•Carrier aggregation •Advanced MIMO techniques •Wireless relaying •eICIC •CoMP transmission/reception •Relay node [HeNB] (femto-cell)

I-2-4 LTE Release 11 to 13

LTE Advanced standard also defines the eMBMS (evolved Multimedia Broadcast / Multicast Service) architecture. It allows greater efficiency in the call service MCPTT for the transmission of voice streams to all the group participants. A new mobile category, category 0, is introduced, allowing lower data throughput and reduced energy consumption. with allowing 88MIMO AND The modulation scheme has been increased from 64-QAM to 256-QAM to increase the downlink throughput

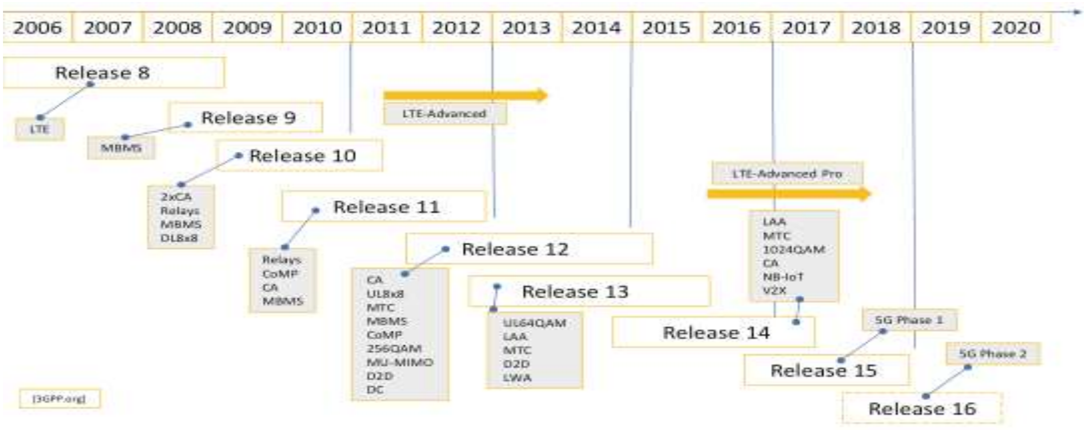


Figure I-3: 3GPP release feature timeline.

Figure I-4 reveals a departure from the typical cellular system evolution which has been given to increasing data rates, increasing user capacity, and lowering latency. This new trend clearly shows the additions of new services (or features, use cases) that the industry is recognizing are required as society demands. These new services were not really intended to be addressed when 4G was created in 2006. The width of new expected services is growing very rapidly .The network is also experiencing its own evolutionary growth.[3]

I-2-5 Release 14

LTE is continuously evolving, addressing not only mobile broadband, but also new areas and use cases. In light of the advancements, Release 13 and onwards is also referred to as "LTE-Advanced Pro." The work on Release [4] 14 has started with the target of being finalized in March 2017 At the same time, the work on the future fifth generation (5G) radio access is ongoing in industry and academia as well as in fora such as the International Telecommunication Union (ITU) and Third Generation Partnership Project (3GPP)

Release 14 focusses on the following items:

- Improving the Mission Critical aspects, in particular with the introduction of Video and Data services
- Introducing the Vehicle-to-Everything (V2X) aspects, in particular the Vehicle-to-Vehicle (V2)
- Improving the Cellular Internet of Things (CIoT) aspects, with 2G, 3G and 4G support of Machine-Type of Communications (MTC)
- Improving the radio interface, in particular by enhancing the aspects related to coordination with WLAN and unlicensed spectrum

- A set of uncorrelated improvements, e.g. on Voice over LTE (VoLTE), IMS, Location reporting.

I-2-6 5G NR phase one : Release 15

Defining an entire new standard for 5G is a large undertaking. 3GPP has split the 5G standard into two releases: Release 15, which corresponds to NR Phase 1, and Release 16, which corresponds to NR Phase 2. In NR Phase 1, there are common elements between LTE and NR, such as both using orthogonal frequency division multiplexing (OFDM). To truly implement the full version of NR, a massive amount of new hardware must be deployed. To continue using existing hardware, a phased approach has been proposed. There is a non-standalone (NSA) version that will use the LTE core and a standalone (SA) version that will use an NR core and be completely independent of the LTE core network.

To keep straight which devices can communicate with each other, new terminology has been introduced:

- LTE eNB : Device that can connect to the EPC or the current LTE core network
- eLTE eNB : Evolution of the LTE eNB that can connect to the EPC and NextGen core
- gNB : 5G NR equivalent of the LTE eNB
- NG : Interface between the NextGen core and the gNB
- NG2: Control plane interface between core network and RAN (S1-C in LTE)
- NG3: User plane interface between the core network and RAN (S1-U in LTE)

Keeping this terminology in mind, the three diagrams from 3GPP TR 38.804 (draft v0.4) shown in Figure 4 and Figure 5 illustrate various deployment scenarios for 5G NR. [11]

The differences between bands are more pronounced for NR due to the very wide range of frequency bands. For NR operation in the new mm-Wave bands above 24 GHz, both devices and base stations will be implemented with partly novel technology and there will be a more widespread use of massive MIMO, beam forming, and highly integrated advanced antenna systems. This creates differences in how RF requirements are defined, how they are measured for performance assessment and ultimately also what the limits for the requirements are set.

Frequency bands within the scope of the present Release 15 work in 3GPP are for this reason divided into two frequency ranges:

• Frequency range 1 (FR1) includes all existing and new bands below 6 GHz.

• Frequency range 2 (FR2) includes new bands in the range 24.25-52.6 GHz.

The frequency bands where NR will operate are in both paired and unpaired spectra, requiring flexibility in the duplex arrangement. For this reason, NR supports both FDD and TDD operation. Some ranges are also defined for SDL or SUL[28].

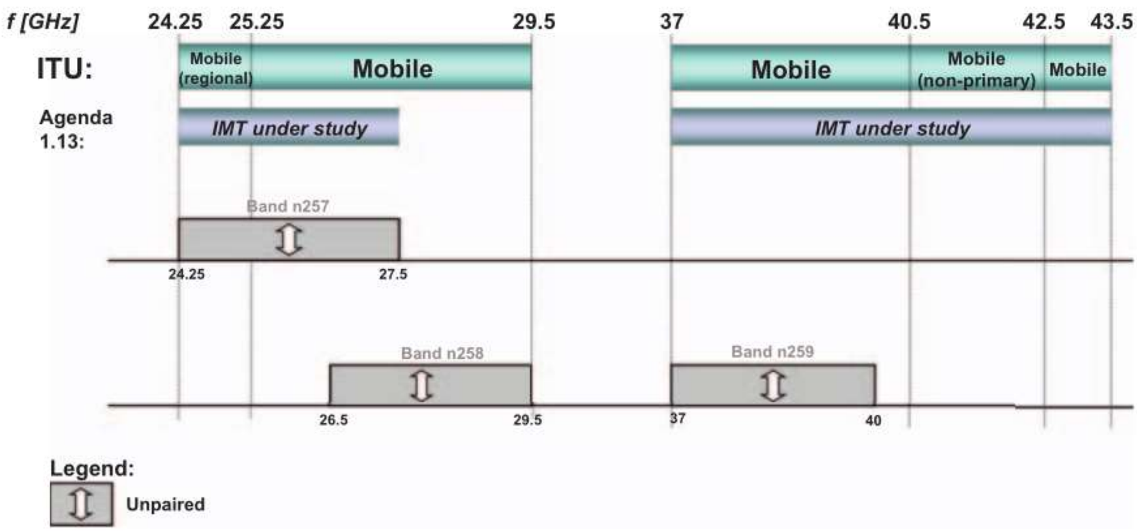


Figure I-4: Operating bands specified in 3GPP release 15 for mmwaves

Release 15 will be completed in the middle of 2018. Significantly reduced latency through the so-called sTTI feature, as well as communication using aerials are two examples of enhancements in this release.

Release 15 Compared to LTE, NR provides many benefits. Some of the main ones are:

• in release 15 ,the new concept of eMBB is introduced as one of they key requirements for 5G .

• the introduction of a new frame structure and spectrum

• exploitation of much higher-frequency bands as a mean to obtain additional spectra to support very wide transmission bandwidths and the associated high data rates;

• ultra-lean design to enhance network energy performance and reduce interference;

• forward compatibility to prepare for future, yet unknown, use cases and technologies;

• low latency to improve performance and enable new use cases

• a beam-centric design enabling extensive usage of beamforming and a massive number of antenna elements not only for data transmission (which to some extent is possible in LTE) but also for control-plane procedures such as initial access. [6]

The first three can be classified as design principles (or requirements on the design) and will be discussed first, followed by a discussion of the key technology components applied to NR.

I-2-7 5G NR phase two : Release 16 and beyond

The Release 16 timing is targeting to have specifications available in mid-2020.[07]

the physical layer details at the end of 2019 and radio protocols by March 2020, with an ASN.1 freeze in June 2020. The work was slightly delayed from the original timing due to the efforts needed for Release 15 finalization, and at the same time there was a desire to ensure early enough stability for the L1 parameters. This was an issue in Release 15 finalization and prevented from use of an earlier than December 2018 version for commercial implementations.

The most notable enhancements to existing features in release 16 are in the areas of multiple-input, multiple-output (MIMO) and beamforming enhancements, dynamic spectrum sharing (DSS), dual connectivity (DC) and carrier aggregation (CA), and user equipment (UE) power saving.

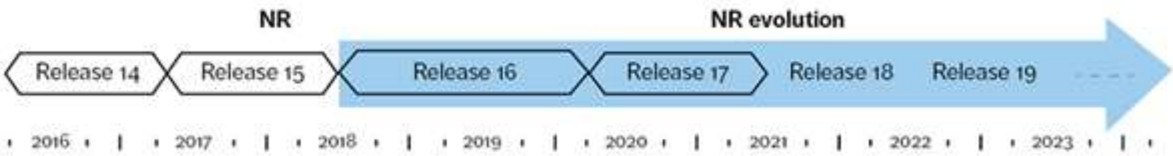


Figure I-5: NR evolution time plan

For release 16 we can summarize its goals in the next points :

• Release 16 introduces enhanced beam handling and channel-state information (CSI) feedback, as well as support for transmission to a single UE from multiple transmission points and full-power transmission from multiple UE antennas in the uplink (UL). Additional mobility enhancements enable reduced handover delays, in particular when applied to beam-management mechanisms used for deployments in millimeter (mm) wave bands .

• In release 16, the number of rate-matching patterns available in NR has been increased to allow spectrum sharing when CA is used for LTE.

• Features related to Industrial Internet of Things (IIoT) and ultra-reliable low latency communication (URLLC)

•Intelligent transportation systems (ITS) and vehicle-to-anything (V2X) communications[08]

• For eMTC and 5G, the PRB cannot be perfectly aligned due to the presence of a DC subcarrier in eMTC. As a result, for 15kHz subcarrier spacings (SCSs), an additional 5G PRB will be needed to accommodate eMTC. Several potential enhancements can be considered, for example, by puncturing the last subcarrier in eMTC so that only six 5G PRBs would be needed. Note that this may be done via implementation with the eMTC performance impact considered.

The contents of 3GPP Rel-17 have already been agreed by 3GPP stakeholders throughout 2019. By June 2021, the Rel-17 standards are expected to be finalized and published.

For release 17 main 16 we can summarize its goals in the next points :

• 3GPP Rel-17 will bring more use cases where mobile communication can be utilized. [09]

• NR Machine Type Communication: In 3GPP, 5G Low Power Wide Area (LPWA) IoT communication is based on the LTE massive MTC (mMTC) technologies: LTE-M and NB-IoT.

• IAB enhancements: We expect the Integrated Access and Backhauling (IAB) solution, which was introduced in Rel-16, will be further evolved to provide increased efficiency and support additional use cases.

• Sidelink enhancements: Rel-17 is expected to support more use cases for 3GPP based mobile communication.

• NR operation on high frequencies (NR > 52.6 GHz): Operating NR in higher than 52.6 GHz provides great capacity for the operators in specific use cases such as mobile broadband services in indoor and dense urban scenarios.

The next steps beyond Release 16 include starting the work on frequency bands higher than 52.6GHz. The next step is under preparation with channel modeling work already under way to cover frequency ranges as high as 114 GHz. Release 17 is expected to cover the study phase for above 52.6GHz, including waveform and other L1 considerations due to higher frequency operation, and in that case the normative work would be left to alter phase of Release 17 or to Release18.Support for higher frequency bands will then also enable 5G-U (NR-U) operation using frequencies higher than 5GHz unlicensed operation, covering especially the 60GHz frequency band. Further, there are more use Cases being worked on in 3GPP from the requirements point of view ,and areas like public safety are expected to be worked on after Release 16. IIoT/URLLC work is expected [10]

I-3 Network architecture

Mobile networks have evolved into a huge multi-radio access technology (multi-RAT) and multi-layer heterogeneous network. But in face of emerging mobile internet applications and digital floods, this architecture becomes more and more incompetent. [5], are gradually introduced to cellular networks. Network architecture based on cloud RAN (C-RAN) and SDN attracts both academic and industrial great attention. According to the above researches, a novel network architecture is presented in Figure I-5. The architecture consists of application cloud, SDN controller cloud, SDN- based C-RAN, SDN-based transport network, and SDN-based core network Application cloud provides various services, such as network management and performance monitor etc. SDN controller cloud transforms policies from application cloud and provides centralized control services for objective network elements by these policies, such as C-RAN, transport network, and core netwrok. The SDN-based C-RAN consists of large-scale baseband unit (BBU) pools, radio-over-fiber (RoF) systems , and distributed radio access points (RAPs) and light RAPs (LRAPs). The BBU pools provide centralized baseband signal processing far beyond single base station. RAPs achieve signaling coverage like macro cell, while LRAPs achieve data transmission like small cell, The RAPs and LRAPs connect BBU via the RoF system. The RoF system achieves intelligent connections between RAPs/LRAPs and BBU. The SDN controller provides dynamical bandwidth adjustment of each RAP/LRAP to BBU connection and BBU pools management including network RAT and version. The SDN-based Transport

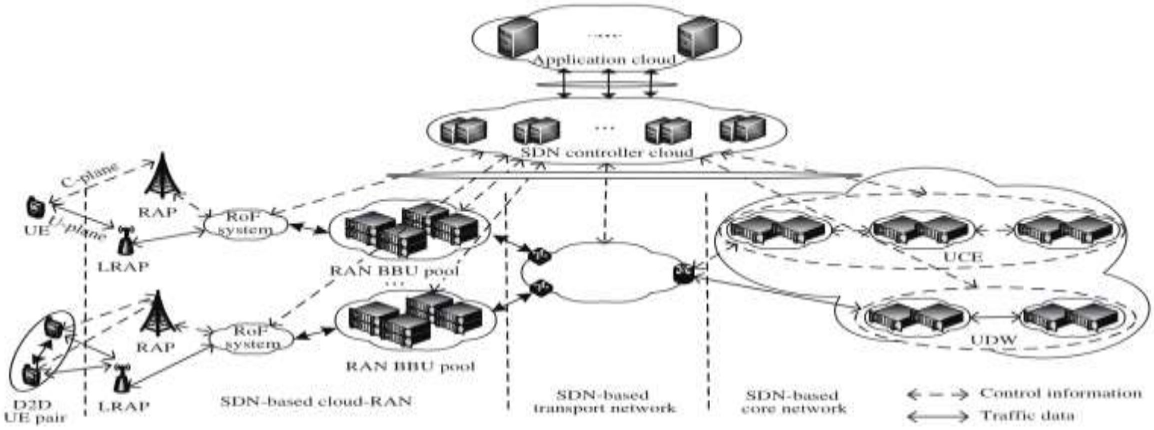


Figure I-6: Network architecture.

SDN-based core network consists of unified control entity (UCE) and unified data gateway (UDW) UCE together with SDN controller manages general packet radio service (GPRS) tunneling protocol for user plane. The cloud technique has several benefits as follows:

1-The large-scaly centralized deployment provides processing capacity far beyond single deployment.

2-Virtualization enables the logical separation of software and hardware, which reduces complexity of processing and expands capacity of hardware

3-New service solutions to dense network deployment, such as radio access network as a service (RANaaS) .In the architecture, application server, SDN controller, RAN and core network adopt the cloud deployment. [12]

SDN decouples network control from individual network devices and migrates it into accessible computing devices Therefore, the SDN can quickly respond to changing network devices, business needs or user demands etc. The SDN architecture consists of application layer, control layer, and infrastructure layer. The interface between application layer and control layer is open application programming interfaces (APIs).

I-4 Key enabling techniques

I-4-1 Non- and quasi-orthogonal multiple access

The quest to improve the spectral efficiency has been regarded as the most important but yet challenging task in the design of future wireless communication systems, due to the fact that the rapid growth of multimedia services, such as interactive game and television applications, cannot be coped with the scarce radio frequency (RF) spectrum resources. Non-orthogonal multiple access (NOMA) [5] and sparse code multiple access (SCMA) are the latest two members of the wireless multiple access technique family to meet future demand for mobile broadband spectrum. NOMA multiplexes multiple users in the power-domain, while SCMA exploits sparse codes to achieve that purpose.

I-4-2 Massive MIMO

Traditional MIMO has been studied extensively in 3G and 4G systems. It is congenital to believe that the capacity would increase linearly with the scale of antennas array if the number of antennas increases largely, i.e., orders of magnitude higher than current configuration, even the simple zero-forcing (ZF) detector would work well, which leads to the so-called massive MIMO ,Next generation mobile communication systems would benefit from massive MIMO as follows:

• Capacity: Let n_t and n_r be the number of transmitting antennas and receiving antennas, and γ be the signal-to-noise ratio. The capacity of MIMO can be bounded by

$$\log_2(1+\gamma n_r) \leq C \leq \min(n_t, n_r).\log_2\left(1+\frac{\gamma \max(n_t,n_r)}{n_t}\right). \qquad \text{(I-1)}$$

In a cellular system, it is likely that the base stations are equipped with massive MIMO antennas while the user equipments (UEs) only have a limited number of antennas, e.g. less than 8. From (1), the capacity would increase dramatically in both uplink (UL) and downlink (DL) transmission.

• Latency: The latency of wireless data transmission is deeply affected by fading, because the retransmission or low rate error control coding would be adopted to resist fading. By employing massive MIMO, the receiving signal would benefit from a large number of space diversity and MIMO signal processing, such as beamforming and precoding, which could compensate the fading.

• Cost and power: Although the number of antennas of massive MIMO is up to the hundreds, massive MIMO can obtain a higher gain with a much lower emitting power per antenna. Actually, the total power of massive MIMO is even much less than traditional MIMO, which means that the low-cost and low-power amplifiers with milli-Watt emitted power would replace the traditional much more expensive ultra-linear amplifiers with tens of Watt power, but there will be still a long way to go before it is used. Some unsolvable problems and future directions include:

• Advanced signal processing algorithm: Coordinating the hundreds antennas to form the beamforming signal is not an easy task. On the other hand, although the emitted power level is lowered by massive MIMO, the power consumption of baseband signal process is increased for more complex process. It is required that signal process algorithms should be simple and effective. Some linear and nearly linear algorithms which could be processed in real time have been proposed . And the tradeoff between complexity and performance should be optimized.

• Channel estimation: The channel estimation could be accomplished by each UE's pilots in uplink. However, the estimation for the downlink is much more sophisticated. It is required that the down link has to have the same number of orthogonal pilots according to the number of hundreds antennas which would cause the so-called pilots contamination. The possible ways to work this out are optimization of pilots allocation, etc.

• Hardware implementations: Although each single antenna unit is simplified and low-cost, its capability to resist phase noise and I/Q imbalance should not be reduced. Especially in high-order modulation, phase noise caused by non-linearity of amplifier will dominate the impairments instead of additive white gaussian noise (AWGN). It is necessary to design sophisticated phase noise compensation algorithms.

• Deployment Scenarios: 5G must face many new deployment scenarios, such as density deployment, D2D communications, relaying coordinated multi-point (CoMP) transmission and reception, etc. How to employ massive MIMO technique in these scenarios effectively still needs further research. [5]

I-4-3 Full duplex

In 4G systems, both frequency division duplex (FDD) and time division duplex (TDD) require two separate channels to realize orthogonal transmission and reception, which wastes half of radio resources. Full-duplex can double spectrum efficiency by simultaneous transmission and reception on the same frequency and time resource. In addition, full-duplex also helps to reduce E2E packet delay and improve network efficiency in contention networks by mitigating the hidden terminal problem. Due to double spectrum efficiency, full-duplex is widely considered as one of the promising techniques in 5G systems.

To promote full-duplex research and application, EU FP7 launched DUPLO project. However, there are still some technical obstacles to be overcome before full-duplex is really put into practice. [13]

• Self-interference cancellation and suppression: Because radio signals attenuate exponentially over the distance, signals from local transmission antennas can be hundreds or thousands of times stronger than those from other BSs.. Hence, Antenna cancellation uses two transmit antennas and one receive antenna with proper distance to overcome self-interference.. Combining RF with digital interference cancellation, antenna cancellation allows full duplex operation. There are several self-interference cancellation and suppression techniques, which are summarized in Table I-2. Current studies mainly adopt two methods: passive suppression and active cancellation. Self-interference processing gains are mainly obtained from antenna and RF stages. Real-time online automated tuning is a challenge. Performance of self-interference suppression and cancellation of wideband signals with high transmit power is still undesirable. High performance digital cancellation solutions still need further investigation.

• Performance analysis and verification: Rate gain region and design trade-off for single antenna, achievable rates under limited dynamic range and fast fading channels with imperfect

Table I-2: Summary of self-interference cancellation and suppression techniques.

Solution	Technical features	Main evaluation	Shortages
Antenna cancellation	Two Tx and one Rx antenna with proper distance between antennas	About 30 dB	Two Tx + one Rx; bandwidth constraint; manually tuning the phase and amplitude.
RF cancellation	- Recreate the RF SI signal by Tx digital baseband signal - Using a circulator, a variable attenuator and phase shifter	-Over 30 dB -About 75 dB	-Extra components to recreate RF SI signal.

	with a general antenna - Convex reformulation for tuning attenuation and phase shift parameters of multiple-tag analog SI canceller - Balanced/unbalanced (Balun) transformer	-About 40–65 dB -At least 45 dB	-Gains vary over bandwidth; -Exact delay, attenuation and phase shift estimation -One Tx + one Rx ; frequency retuning; high insertion loss
Antenna + suppression	-Directional separation, RF absorptive shielding and cross-polarization antenna	-Over 70 dB	not effective in reflection path; FR selectivity
Antenna suppression + RF suppression	-Dual-polarized dual-feed antenna and electrical balance duplexer	Both over 50 dB, size of duplexer and antenna is small	High Tx insertion loss (3.4 dB) and cascaded Rx noise (7 dB)

Channel estimation [5], Precoding and degrees of freedom for multi-antenna, precoding and propagation methods for multi-users and peak performance improvements in multi-cell scenarios have been studied.. The analyses for complex scenarios, such as massive MIMO and large scale multi-cell multi-user field verification, are still lack. Meanwhile, as FD is more suitable for short range communication, research on small cell scenarios with multiple users may attract more attention in the future.

I-4-4 Small cells

Small cells are smaller and cheaper than a cell tower. They can be installed in a variety of areas., bringing more base stations closer to users. A large number of base stations increases the number of people a network can support, increasing the network density and improving the throughput is to densify the number of wireless nodes, which have a smaller coverage range than the macro-cell base stations used in the 3G and 4G legacy systems. The technical solution behind this idea is denoted as the small cell technology. The “small cells” is an umbrella term for operator-controlled, low-powered radio access nodes with a coverage range in between ten to several hundreds of meters. including licensed and unlicensed (WiFi) . An example of small cell deployment is shown In Figure I-6 With small cells, the size of the cell is reduced, meaning the network is closer to the user.. In addition, The higher number of low-powered transmission points enables better use of available frequency resource. Furthermore, the 5G system will be constructed in a heterogeneous fashion, where macro and small cells are co - located or connected

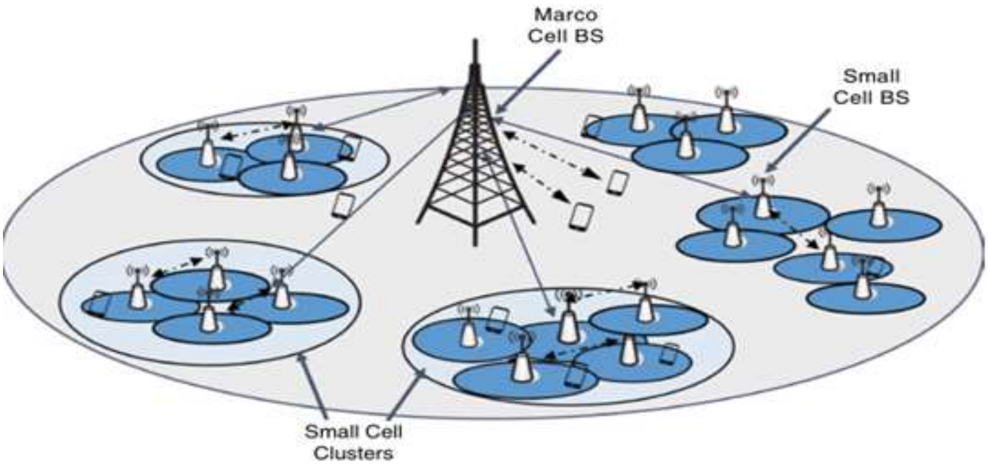


Figure I-7: An illustration of small cells deployment.

via wireless backhaul links, thus providing increased levels of network capacity through traffic offloading. However, the heterogeneity of small cells in the network will pose challenges in terms of interference and mobility management. These issues will need to be addressed in future research. Some other ongoing research on small cells for 5G would include load balancing, wireless backhauling, mmWave and massive MIMO in small cells, etc. [14]

I-4-5 Device-to-device communications

Proximity-based D2D communications underlying cellular networks are a highly efficient way to enhance system capacity. Besides, D2D UEs (DUEs) can act as transmission relays for each other to set up multi-hop communication links. Therefore, it also helps to improve and extend network coverage by DUE relaying. The gain of D2D communications depends on the number of available DUE pairs in various application scenarios. The standardization of this technique is on-going in 3GPP and will be available in Release 12, In 5G era, with the wide use of various smart terminal devices, the communication range between users will be shortened, To apply this new technique to 5G systems, the following issues need to be addressed at least:

• Direct discovery: Devices must know their neighbors before directly communicating with each other. proximity-based device discovering and services. Direct discovery includes two models: model A (“I’m here”) and model B (“Who is here”/“are you here”) .In model A, discoveree UE announces its existence with certain information about itself; discoverer UE will read and process the information only if it is interested. In model B, dis- coverer UE transmits a request with certain information about what it is interested to discover; discoveree UE will respond if it meets such a request. [15].

• Interference management: As D2D communications share resource in cellular networks, it inevitably generates interference among UEs .Various interference scenarios in full-duplex D2D communications are summarized in Figure I-7 D2D pairs should keep a certain distance away from BS and primary cellular UEs.

• Direct communication: In 4G systems, physical channels of direct communication link reuse physical uplink shared channel (PUSCH) structures . As single carrier frequency division multiple access (SC- FDMA) has low peak-to-average power ratio (PAPR), it reduces the dynamic range requirements of power amplifier (PA) and improves power efficiency. While in 5G systems, the available frequency resources may disperse in several frequency bands. To flexibly utilize these dispersed frequency resources, new multi-carrier technologies like filter bank multi-carrier (FBMC) are considered. How to design D2D direct communication link is still a fresh topic.

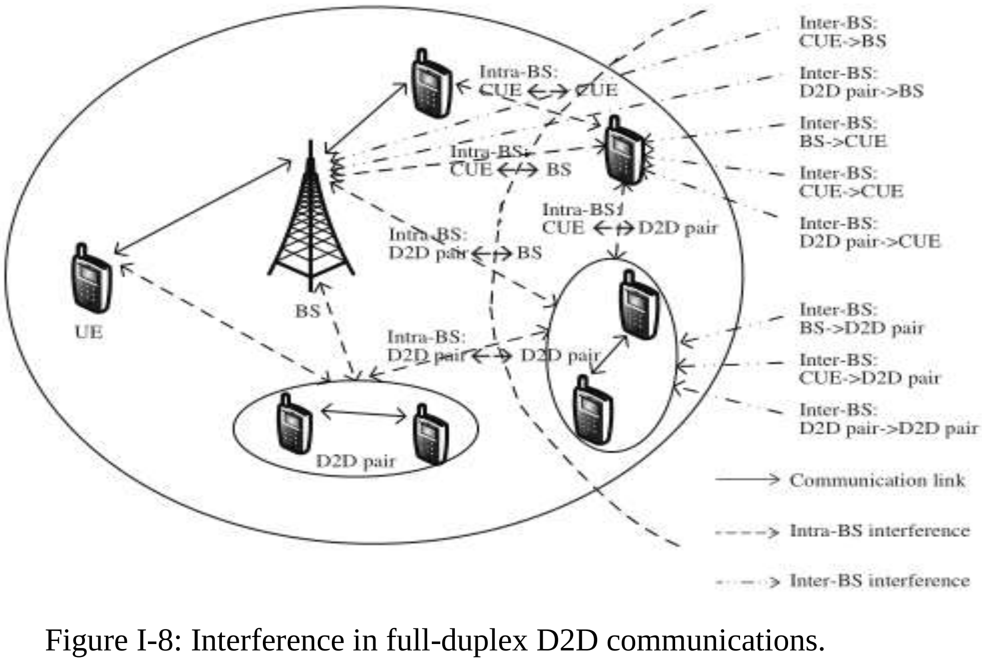


Figure I-8: Interference in full-duplex D2D communications.

I-4-6 Millimeter wave communications

One of the efficient ways to satisfy rapid increase of data rates, especially those up to tens of Gbps in 5G systems, is bandwidth expansion. Most of the cellular networks work below 3 GHz which has been fully occupied already. Bandwidth shortage has motivated the exploration of the rich millimeter wave (mmWave) frequency spectrum which ranges from 3 to 300 GHz. There are potential dozens of GHz available frequency resources at 28, 38, 45, and 60 GHz. It is expected that the gains of network capacity up to 10 times can be obtained from mmWave frequency spectrum which is quite attractive for 5G systems.

Table I-3: Atmospheric attenuation windows.

Frequency (GHz)	3	23	31	60	78	119	127	183	214
Atmospheric attenuation (dB/km)	0.0075106	0.19488	0.10003	15.17285	0.35743	2.04379	0.86255	28.36202	2.72848

There are already on-going academic and industrial efforts to study the feasibility of mmWave communications. EU FP7 launched the MiwaveS project, facing beyond 2020 heterogeneous wireless networks with mmWave small cell access and backhauling. Samsung has demonstrated over 1Gbps download rate at 28 GHz frequency band. Although some progress in mmWave communications has been made, there is still a long way before realizing practical 5G mmWave communications. The main technical obstacles include the following aspects[5]:

• Propagation characteristics and measurements: Radio waves of different frequency band have different propagation characteristics. Traditional cellular wireless communication channel models. The first task of mmWave communications is to understand the propagation, as shown in Figure I-11, the atmospheric attenuation is 0.01–40 dB/km at mmWave frequency, which is much higher than 0.001–40 dB/km of frequency bands used by traditional cellular networks. Moreover, there are five atmospheric attenuation windows, which is shown in Table

I-4. rain attenuation up to 0.001–40dB/km varies with frequency and rain rate at mmWave frequency., which is much higher than 0–0.001 dB/km level of traditional used frequency bands; the reflection coefficients are up to 0.896 outdoor and 0.74 indoor separately, which are still higher than traditional used frequency bands, which is shown in Table 4. The path loss exponent is slightly higher than 2 for line-of-sight (LOS) channels and 4 for nonlinear-of-sight (NLOS), which is shown in Table I-5, but in some specific environments, such as in vehicles, the path loss exponent can be up to 7. The root mean square (RMS) delay spread of LOS has little change and the RMS delay spread of NLOS decreases with frequency, which is shown in Table I-6; besides, the RMS delay spread of NLOS also decreases with distance. Table I-8 shows that mmWave communications suffer from severe outage when distance is beyond 200 m, The effective communication distance of mmWave signals is within 200 m..

• Antenna arrays: Because of the unfavorable mmWave propagation characteristics, it can be used for short range communication in 5G systems. Antenna arrays are considered as a key technique to achieve mmWave communications, Large antenna arrays with beamforming can provide sufficient gains to overcome the severe propagation loss. Besides, it also needs to adapt promptly when beams are blocked. Hence, high efficiency low complexity adaptive antenna array need to be studied

• Hardware implementation: One of the major challenges comes from the high power consumption of mixed signal components, such as the analog-to-digital converters (ADCs) and digital-to-analog converters (DACs). The wideband mmWave signals up to 1 GHz are far more than bandwidth of current ADCs/DACs. Besides, increasing transmit power requires highly linear and efficient PAs. Large array antennas also need innovative hardware architecture of transceiver.

Table I-4: Reflection coefficients at 28 GHz.

Environment	Location	Material	Angle (◦)	Reflection coefficient
Outdoor	ORH	-Tinted Glass -Concrete	10 10 45	0.896 0.815 0.623
Indoor	MTC	- Clear Glass – Drywall	10 10	0.740

			45	0.704 0.628

Table I-5: Path loss exponents.

Frequency (GHz)	28	38	40	60	72 2.55
Path loss exponents (LOS)	2.55	2.0	—	0.8	—
Path loss exponents (NLOS)	5.76	4.57	—	7.4	—

Table I-6: RMS delay spreads.

Frequency (GHz)	28	38	40	60	72
-RMS delay spread (LOS) (μs)	0.878	1.2	—	0.8	—
-RMS delay spread (NLOS) (μs	47.2	23.6	—	7.4	—

Table I-7: Outage statistic at 38 GHz .

Tx location	Height (m)	Outage (%) for 160 dB PL sensitivity	Outage (%) for 150 dB PL sensitivity
TX1 ENS	36	18.9% within 400 m, 0% within 200 m	52.8% within 400 m, 27.3% within 200 m
TX2 WRW	18	39.6% within 400 m, 0% within 200 m	52.8% within 400 m, 10% within 200 m

I-4-7 Beamforming

Beamforming is a method to concentrate the omni-directional radiated power of antenna. In 5G system beamforming is referred to transmission of a signal in a narrower shape from base station to receiver. While transmitting, main lobe of the antenna pattern is directed towards a certain direction using phase and amplitude of each antenna element in an array. This phenomenon is called constructive interference in the wave front. Figure I-8 depicts a typical beamforming scenario in case of cellular technology.

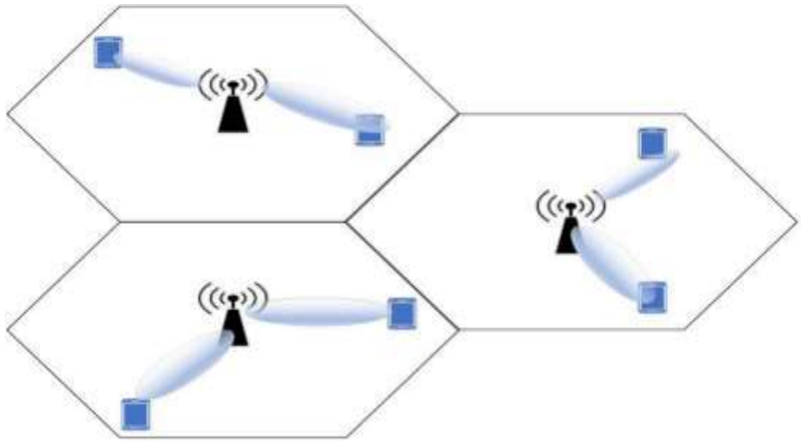


Figure I-9: Typical beamforming scenario inside a cell.

The Beamforming uses multiple antennas to generate/receive electromagnetic wave with controllable beam pattern. As shown in Figure I-9[16], to make specific slanted RF signals (ex:20 degree), the feeding signals of each antenna have their own designated signal delay which is proportional to antenna space and slanted angle. The relation among them

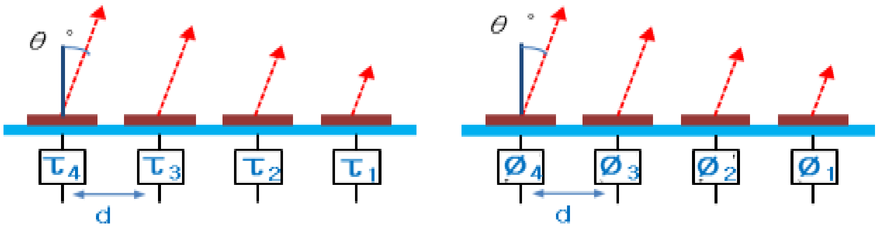


Figure I-10: Beamforming architecture: (a) ideal Beamforming (b) phase shift equivalent Beamforming.

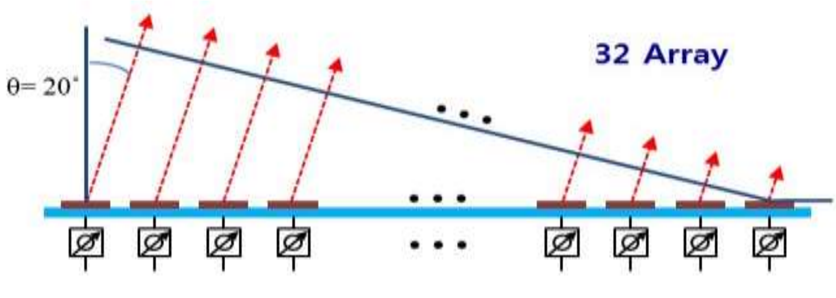


Figure I-11: Beam Squint example with linear 32 antenna elements and 200 beam direction.

Beamforming can be done on both transmitter side and receiver side. Beamforming is a wide concept as it is associated with beamforming management and beam training , Millimeter waves are susceptible to path loss in free-space, rain, fog and other atmospheric events , Figure I-11 was generated using Friis' free-space path loss model [17] in order to compare free-space path loss in 5G and 4G. Two frequencies were chosen for 5G (28 GHz and 300GHz) and one for 4G (2100 MHz). If the distance between the transmitter and receiver is d (km) and the frequency is f (GHz) then the free-space path loss (FSPL) in dB) can be calculated as

$$FSPL = 92.4 + 20\log(f\ in\ GHz). \quad \text{(I-2)}$$

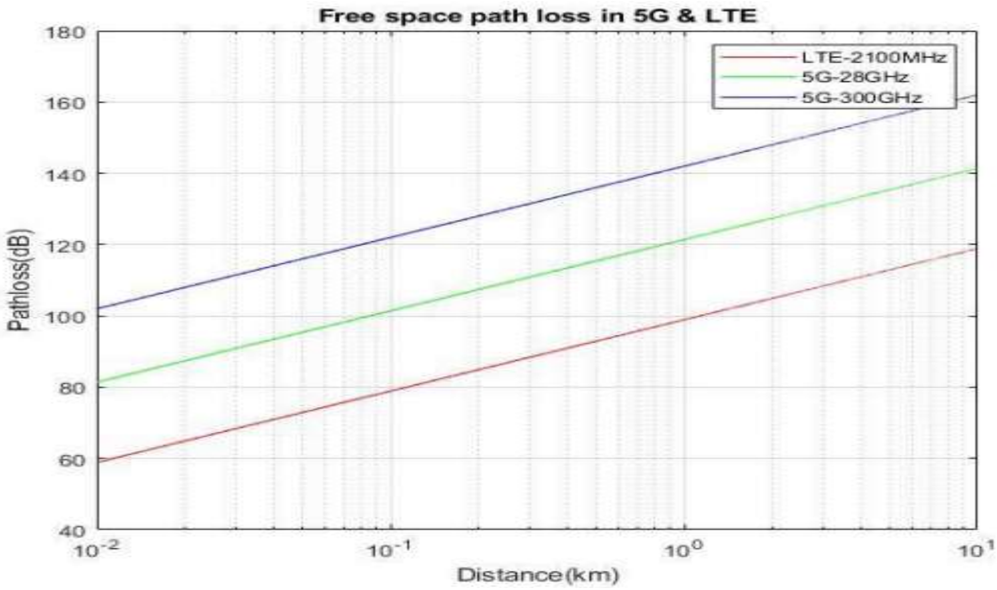


Figure I-12**:** Free-space path loss for different frequencies.

From I-11 it is visible that for a given range of distances (10m-10km), corresponding path loss increases (approx. 20 dB) with the higher frequency. To overcome this high path loss in 5G mmWave, beamforming is the ultimate way to increase the directivity of the antenna. Antenna array of different sizes are used to minimize path loss. Different kinds of arrays are used to exploit the best antenna geometry. Mostly used arrays are uniform linear array (ULA), uniform rectangular array (URA) and uniform circular array (UCA). The number of antenna elements used in array defines the array gain . Each antenna element in an array is separated by half of the wavelength of the transmitted frequency. The number of antenna elements used in array defines

I-4-8 Green communications

Large scale mobile communication networks have become a non-negligible part of the world energy consumption, China Mobile has proposed two major themes for 5G systems. One of the two is green technology . One of the three efficiency indexes of 5G systems proposed by IMT-2020(5G) PROMOTION GROUP is energy efficiency. BSs account for more than 50%

of energy consumption in telecommunication systems. Researches on green communications mainly focus on the following aspects [5]:

• Improved network architecture and deployment: Traditional cellular-based designs mainly focus on seamless coverage and system capacity issues. Green design of cellular networks is becoming a major issue. Control/user plane separation and heterogeneous networks are considered as promising architecture.

• Green radio resource management: Radio resource management (RRM) is an important part of the RAN and various algorithms have been studied Traditional RRM mainly concentrates on spectrum efficiency and QoS. RRM also plays an important role in reducing energy consumption Green RRM requires integrated tradeoffs, such as deployment efficiency and energy efficiency. Game theory based resource allocation algorithm has received special attention. More efforts should be devoted to the design of unified green RRM frameworks.

• Energy efficient power amplifier: Power amplifier is the biggest energy consumption component in BSs. It can consume up to 22% of all energy consumption.. 5G systems, data rate up to 10 Gbps requires much wider power range. Current PA techniques can hardly satisfy such requirements. It is necessary to put emphasis on highly linear and efficient wideband multi-frequency band PA.

• Sleeping for BSs: Due to non-uniform traffic distribution, it's a waste of energy to keep all BSs working around the clock. In the period of low traffic, only a fraction of BSs need to provide the service, and other BSs may enter low-power sleeping mode to reduce the power consumption. Several sleeping algorithms have been proposed . These approaches need to pre-configure coverage cells to guarantee coverage, which adds to configuration complexity. In 5G era, it will be an ultra-dense and heterogeneous network, and the problems of how to efficiently select sleeping BSs are still open.

I-5 Architecture of 5G

Non-standalone and Standalone: two standards-based paths to 5GAs with the previous generations, 3GPP is defining both a new 5G core network, referred to as 5GC, as well as a new radio access technology called 5G "New Radio" (NR). Unlike previous generations that required that both access and core network of the same generation to be deployed (e.g. Evolved Packet Core (EPC) and LTE together formed a 4G system), with 5G it is possible to integrate elements of different generations in different configurations, namely : [18]

• Standalone using: only one radio access technology

•Non-Standalone: could use multiple radio access technologies.

I-5-1 Standalone (SA) network architecture

The first network deployment mode is referred to as standalone. SA refers to having an independent 5G network. 3GPP finalized the standalone 5G NR standard in 2018, which will work alongside the Non-Standalone 5G NR standard. Standalone 5G NR will have a new end-to-end architecture that will use mm-Waves and sub-GHz frequencies. This mode will not use existing 4G/LTE infrastructure.

Standalone 5G NR will use enhanced mobile broadband (eMBB), Ultra-reliable and low latency communications (URLLC) and Massive machine type communications (mMTC) to provide multi-gigabit data rates with improved efficiency and lower costs. It will have both the new 5G air interface, New Radio (NR), and 5G Core (5GC) in place. A standalone 5G network provides the user an end-to-end 5G experience. The SA network will still interoperate with the existing 4G/LTE network to provide service continuity between the two network generations.

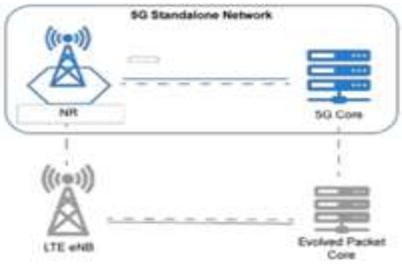


Figure I-13: 5G SA network architecture.

As shown in the figure 11, the 5G network can operate independently. At the same time, interoperation with LTE network takes place in order to cover areas not yet covered by 5G and to connect 5G users with non-5G users

I-5-2 Non-standalone (NSA) network architecture

Non-standalone 5G network, on the other hand, refers to having only 5G NR cells in place with EPC as the core.[18] Operators will deploy 5G cells and depend entirely on existing LTE network for all control functions and add-on services. The 5G NSA architecture works in master-slave structure, where the 4G access node is the master and 5G access node is the slave. utilizing a 4G EPC with a 5G NodeB (gNB). This option is most popular for mobile operators looking to quickly deploy 5G speeds utilizing existing LTE deployments. However, NSA

Option 3 does not allow for true 5G NR features, such as Network Slicing, URLLC and high capacity support for IoT, such as mMTC.

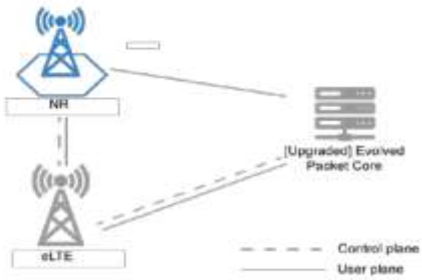


Figure I-14: 5G NSA network architecture.

I-6 Conclusion

The road map for 5G have seen many architectures ,the acceleration of deployment of the 5th generation is done in the release 15 where for the first time in 3GPP history they made two phases ,NSA which introduces the air interface of the 5G where the Data plane is forwarded to the EPC and still the control plane done by the 4th Generation core network , this implementation is adopted by many countries around the world as the fist step to 5G , Most of the countries in the world, such as Japan, China, and those in Europe, have taken significant steps to lead global developments toward this technology. Mobile communication is undergoing a revolutionary shift over the world because of the increase in the number of mobile users. Additionally, consumers will be able to download a 1080p high definition (HD) movie to their mobile phone in about one second and 50 GB video games within minutes with this 5G mobile communication technology.the eMBB application could fulfill all the needs of the high data rate ,as well as NSA, 5G mobile communication is the next step in the transformation of Communication technology .The competence of 5G mobile communication must pass far beyond those of the previous generations to facilitate connectivity for a wide range of applications. For the future releases the SA will be implemented and a new applications would be supported , in this chapter We provided an overview of several emerging technologies for future 5G mobile communication networks. In addition, to achieve future 5G technology, the technologies such as massive MIMO and beamforming, higher bandwidth at mm-Wave, VLC, D2D communication, advanced multiple access, and modulation will also impact the design and development of 5G networks.

Chapter II: Latency in 5G

II-1 Introduction

Latency vertical industries such as robots and interactive internet tools and computers, as well as the Internet-of - Things (IoT). , In particular, including latency and end-to - end responsiveness (E2E), network robustness, and performance, both of which play a significant role in the level of service required to meet such goals, the next generation network is moving the 5 G standards into various areas such as data rate, latency, efficiency, etc., Energy consumption device / network, traffic volume rate, accessibility and link speed. Present fourth generation (4 G) networks are not capable of satisfying all the technological specifications for such latest technologies.

In order to reach low latency, certain improvements need to be made to the network design, so they do need to be applied. As the largest latency is the connection of the radio access network (RAN) and the core network together with the backhaul between the RAN and the core network, the current network topology comprising software specifying the network (SDN) is mobile edge computing. (MEC)/caching, network virtualized feature (NFV) may be used to greatly minimize latency. It will happen with the potential to benefit with the device technology that has been implemented by the organizations.Moreover, a new physical air-interface of small packets, different waveforms, short time interval delivery, and coding and modern modulation schemes are the essential fields of research to achieve low latency. The optimisation of the distribution of radio resources, such as mMIMO, gives priority to data transmission , carrier aggregation used in millimeter waves need to be discussed, stable and effective convergence with current LTE services is essential for the 5 G networks to begin with, since that would enable the 5 G technologies to operate even quicker and more effectively when integrated and ready for costumer use. As a result, the 5 G broadband communication network should be an evolution of LTE, in same time, complemented with radio technologies and revolutionary architecture.

II-2 QoS Parameters

The quality of various service parameters plays a key role in the design of next generation networks. These applications are data hungry, To support all these high-density services, the guarantee of QoS parameters is a major objective Figure II-1. To fulfill the requirement of various data-hungry applications in the next generation wireless networks, providing QoS guarantees for these applications is most important. These are the QoS parameters which are frequently used in wireless networks [19] :

II-2-1 Bandwidth

Bandwidth is one of the most important QoS parameters necessary in order to fulfill the demand of next generation futuristic applications. Bandwidth is directly proportional to the amount of data transmitted or the capacity of the channel i.e., Radio Access Network capacity). It is measured in bits/s.

II-2-2 Throughput

The success of the network depends upon the value of throughput. A higher throughput represents the maximum number of bits that are successfully transmitted out of the total available number of bits in the entire message per unit time. Hardware devices, bandwidth, signal to noise ratio (SNR), and others things affect the value of throughput [20] Throughput plays a key role in measuring the actual rate of transmission. The entire gamut of applications depends upon higher throughput.

II-2-3 Jitter

Everyone wants continuous transmission of packets without any delay, but jitter is one parameter which hampers the entire communication by introducing a delay between the packets. It is also called packet delay variation (PDV) Seamless transmission of packets depends upon a minimum delay occurring between packets. This factor must be within the permissible range, especially in the case of next generation video and audio streaming. Jitter is one of the important QoS performance parameters.

II-2-4 Bit Error Rate

This is the parameter which identifies the error occurring in the number of bits out of the total available number of bits transferred. It is also known as bit error ratio. It is defined as follows: It is the ratio between the numbers of bit errors to the total number of transferred bits. Bit error rate (BER) must be minimized to fulfill the requirement of next generation data applications.

II-2-5 Packet Delivery Ratio (PDR)

As the name indicates, this is the ratio of the number of packets delivered out of the total number of packets sent across the network. PDR decides the efficiency of the next generation network. It must be high.

II-2-6 Route Discovery Time

The time taken to determine a route from source to destination is known as route discovery time. In the case of a mobile ad hoc network, the network changes frequently due to node movements. The routing algorithm discovers the appropriate route before transmitting. The route discovery time should be minimum to connect different nodes. Smart routing algorithms are needed to achieve minimum route discovery times for next generation applications.

II-2-7 Latency

Delay that occurs in a communication network is known as latency. There are two kinds of delay: low delay and high delay. The delay between the two ends from source to destination is known as an end-to-end delay (EED) [20]

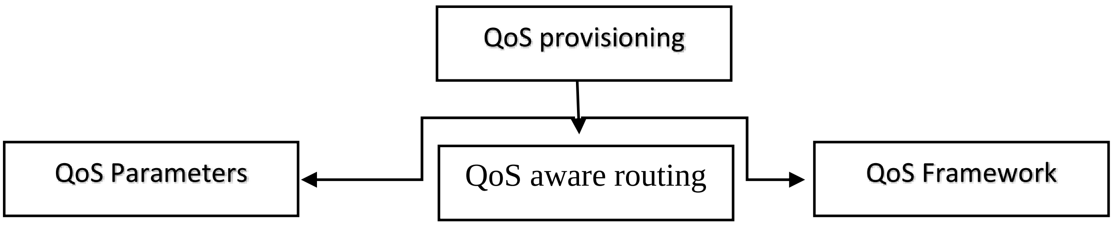


Figure II-1: QoS provisioning.

Latency is dependent upon four components, which are explained by following mathematical equation:[19]

$$\text{Latency} = P.T + T.T + Q.T + P.D. \quad \text{(II-1)}$$

where

- P.T. is propagation time
- T.T. is transmission time
- Q.T.is queuing time
- P.D. is processing delay

In order to achieve a good quality network, latency must be in the minimum range. In our project we will focus on the parameter of latency in the 5G generation:

In order to enable certain services related to Tactile Internet, a minimal level of reliability and latency is required. as the time it takes to transfer a given packet or a part of information from a certain source to a destination , Data is measured from the moment it is transmitted and stamped by the source until the destination

To meet challenging latency requirements, several improvements are applied in so the wireless transmission needs to be shortened significantly. It is proposed to reduce the transmission time interval which is 1 ms in LTE to 0.1-0.25 ms or below . In addition, the waveform has to support short packets and low latency. broadband communication systems which we have today as most are based on OFDM because of the robustness against multi-path channels. However, to achieve low latency transmissions in the order of a millisecond, the physical transmission must have very small packets, which requires a one-way PHY layer transmission of several micro-seconds, the numerology together with the reference symbols design, channel estimation, and channel coding requires significant revision of the cellular PHY for the Tactile Internet applications in 5G systems[20], One way to overcome these limitations for achieving latencies of 1 ms in OFDM-based systems is to change the OFDM numerology, i.e. symbol duration, sub-carrier spacing, etc, together with the sub-frame length, and enable high levels of diversity, and fast channel estimation together with fast channel decoding (e.g, with convolutional codes). it has been shown that an OFDM-based system with changed OFDM numerology can achieve low latency. Further, a recent investigation practically demonstrates achieving 1 ms latency with modified OFDM numerology.

To support the co-existence of vertical applications, while achieving latency of 1 ms, it is desirable to have a flexible frame structure. One approach is to have a dynamic adjustment of TTI in accordance with the service requirements

II-3 Technical Performance Requirements for 5G

Based on the key capabilities and IMT-2020 vision , technical performance requirements are defined in Report ITU-R M.2410 [25] The technical performance requirements are summarized in Table II-1. The detailed definition on the technical performance requirements can be found in To reach the 5G vision defined by ITU-R, 3GPP further studied the deployment scenarios and the related requirements associated with the three usage scenarios as documented in 3GPP TR 38.913 These requirements are usually higher than ITU's technical performance requirement, showing 3GPP's ambition of providing higher capability than ITU required. A detailed description of the 3GPP requirements is beyond the scope of this book. This interested reader is encouraged to consult [24]

Table II-1: URLLC Technical Performance Requirement.

Technical performance requirement	DL	UL	Comparison to IMT-Advanced requirement
User plane latency	1 ms	1 ms	>10x reduction compared to IMT-advanced
Control plane latency	20 ms	20 ms	>5x reduction compared to IMT-advanced
Mobility interruption time	0	0	Much reduced
Reliability	99.9999% within 1 ms	99.9999% within 1 ms	-

II-4 Sources of Latency in a Cellular Network

In the cellular network system for example (LTE), the latency can be divided into two major parts:

- user plane (U-plane) latency
- control plane (C-plane) latency

The measurement of the User plane latency is only by one directional transmit time of a packet to become available in the IP layer between evolved UMTS terrestrial radio access network (E-UTRAN) edge/UE and UE/E-UTRAN node On the other hand, we define the Control plane latency as the transition time of a UE to switch from idle state to active state. At the idle state, an UE is not connected with radio resource control (RRC). After the RRC connection is being setup, the UE switches from idle state into connected state and then enters into active state after moving into dedicated mode, U-plane is the main focus of interest for low latency communication, Since the application performance is dependent mainly on the U-plane latency [22]

The delay of a transmitting a packet in a cellular network can be achieved by combining the RAN, backhaul, core network, and data center/Internet. As referred in Figure II-2

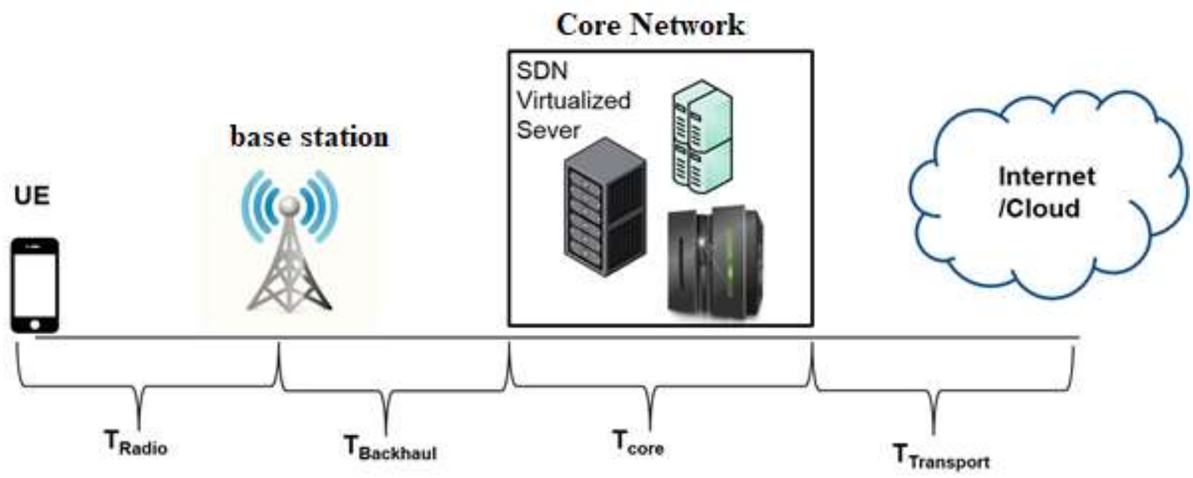


Figure II-2: Latency in E2E delay of packet transmission.

After analyzing the figure, the total one-way transmission time [22] of current cellular system can be written as

$$T = T_{radio} + T_{backhaul} + T_{core} + T_{transport} \quad \text{(II-2)}$$

where

• $\boldsymbol{T_{radio}}$ is the time of transmitting a packet between eNB/gNB and UEs and is mainly due to physical layer communication It is contributed by eNBs/gNBs, UEs and environment. It consists of time to transmit, processing time at eNB/UE, retransmissions, and propagation delay. Processing delay at the eNB involves channel coding, rate matching, scrambling, cyclic redundancy check (CRC) attachment, precoding, modulation mapper, layer mapper, resource element mapper, and OFDM signal generation.

• $\boldsymbol{T_{backhaul}}$ is the time for building connections between eNB/ gNB and the core network (i.e. EPC/CN). Generally, the core network and eNB/gNB are connected by copper wires or microwave or optical fibers. In general, microwave involves lower latency while optic fibers come with comparatively higher latency. However, spectrum limitation may curb the capacity of microwave [22]

• $\boldsymbol{T_{core}}$ is the processing time taken by the core network. It is contributed by various core network entities such as mobility management entity (MME), serving GPRS support node (SGSN), and SDN/NFV. The processing steps of core network includes NAS security, EPS bearer control, mobility anchoring, UE IP address allocation, idle state mobility handling, and packet filtering:

• $\boldsymbol{T_{transport}}$ is the delay to data communication between the core network and Internet/cloud. Generally, distance between the core network and the server, bandwidth, and communication protocol affect this latency

So, The E2E delay, T_{E2E} is then approximately given by 2 x T. The T_{radio} is the sum of transmit time, propagation latency, processing time channel estimation, encoding and decoding time for first time), and retransmission time (due to packet loss). In particular, the $\boldsymbol{T_{radio}}$ for a scheduled user can be expressed as:

$$T_{radio} = T_Q + T_{FA} + T_{tx} + T_{bsp} + T_{mpt} \quad \text{(II-3)}$$

Where:

• $\boldsymbol{T_Q}$ is the queuing delay that depends on the multiplexed number of users that on the same resources.

• $\boldsymbol{T_{FA}}$ is the delay for the alignment of the frame that depends on

Its structure and duplexing modes (FDD mode, TDD mode)

•$\boldsymbol{T_{tx}}$ transmission processing time, and payload transmission which uses at least one TTI depending on radio channel condition, payload size, available resources, transmission errors and retransmission;

• $\boldsymbol{T_{bsp}}$ is the processing delay at the base station;

• $\boldsymbol{T_{mpt}}$ user terminal processing delay of. Both the BS and UE delay, it depends on the capabilities of BS and user respectively.

II-5 URLLC Physical Layer in 5G NR

In contrast to the 4G LTE systems, latency, reliability, and throughput requirements should be jointly considered in 5G NR so that there should be a fundamental change in the physical layer architecture (packet, slot, and frame). Specifically, a latency-sensitive packet structure for the fast decoding process and a flexible frame structure.. Also, when the URLLC service is initiated, the URLLC packet are transmitted instantly with no delay. To achieve that we need to have scheduling scheme that helps for minimizing the transmit latency of the URLLC packet. Hence, since the latency requirement is not satisfied by the HARQ re-transmission, unless TTI of a packet is very short, this type of mechanism that significantly reduces the re-transmission latency is needed.

the physical layer solutions for URLLC including the packet and frame structure to minimize the latency, multiplexing schemes to overlay the URLLC service into enhanced mobile broadband and mMTC services, and approaches to deal with the already existing problems. We note that the reliability improvement and the latency reduction are equally. However, There has been a consensus in the 3GPP NR standard meeting that latency is a priority. [24]

II-5-1 Packet Structure

The key issue in the URLLC packet design is to minimize the processing latency $\boldsymbol{T_{proc}}$ and the time-to-transmit latency $\boldsymbol{T_{itt}}$. Note that $\boldsymbol{T_{proc}}$ consists of the time to receive packets, acquire channel information, extract control (scheduling) information, decode data packets, and check errors. in the 5G NR systems, as we have a non-square packet spread in the axis of frequency, is used as a baseline, the structure minimizes the latency of transmission $\boldsymbol{T_{itt}}$ Furthermore, for latency $\boldsymbol{T_{proc}}$ reducing , the 3 components of a packet (data part ,pilot, control) should be gathered all together to make a pipelined processing of the channel acquisition, data detection, control channel decoding

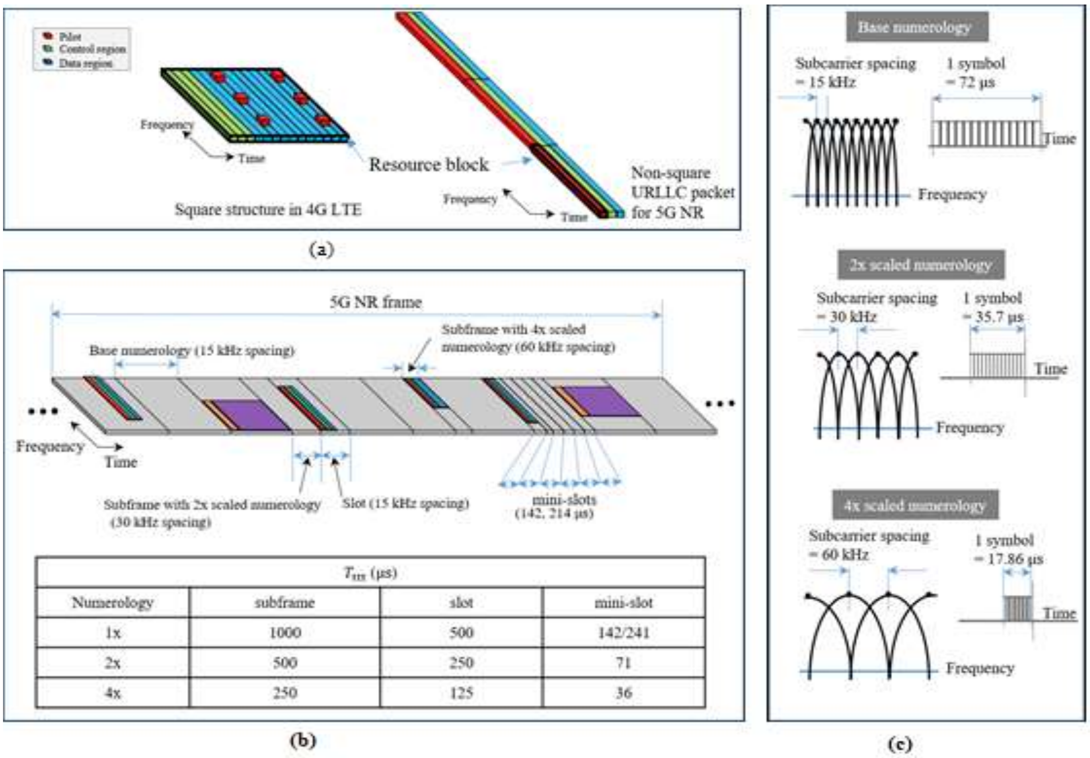
In 5G NR, low density parity check (LDPC) code and Polar code are adopted for enhancing the data and control channel, respectively. Over the years, many efforts have been made to improve the decoding performance and computational complexity (and hence processing latency) of these codes such as successive cancellation list decoding of Polar code and non-binary LDPC decoding [29]

II-5-2 Latency-sensitive Scheduling and Frame Structure

designing a unified frame structure to cover a wide-range of frequency band and various service categories is one of the main goals in 5G NR, to this end flexible frame scheduling mechanism have been introduced.

II-5-2-1 Flexible Frame for URLLC

One direct option to reduce the time-to-transmit latency 7 is to reduce the symbol period (see Figure II-4(b)). When the frequency band above 6 GHz (millimeter wave) , due to the path loss, cell radius would be much smaller than that of conventional cellular systems and so will be the channel delay spread. However, when the frequency band below 6 GHz is used, this option might not be desirable due to the large delay spread. In this case, considering reducing TTI of the packet. As an example, the use of mini slot frame level (3-4 symbols), also slot level (7 symbols) transmission, it means by controlling the number of symbols and the symbol period in a packet, T_{itt} is being smaller than 1ms can be achieved (see Table in Figure II-4(b)). Note that to support this flexible frame structure, an advanced receiver is needed [29]



T_{tts} (μs)			
Numerology	subframe	slot	mini-slot
1x	1000	500	142/241
2x	500	250	71
4x	250	125	36

Figure II-3: Packet and frame structure for URLLC: (a) packet structure; (b) frame structure; (c) supported numerologies for 5G NR.

.

II-5-2-1-1 Instant scheduling scheme

the ongoing data transmission is interrupted to initiate the URLLC packet. This protocol is effective for reducing the URLLC time access but it could cause a noticeable degradation in the performance, Therefore, we need to mitigate to mitigate the performance degradation by having an approach of ongoing services.

II-5-2-1-2 Reservation based scheduling scheme

the reserve of URLLC resources have a priority the data scheduling. Two types of reservation schemes are semi-static and dynamic reservations in the semi-static reservation scheme, the base station infrequently broadcasts the configuration of the frame structure such as frequency numerology and service period. Whereas, in the dynamic reservation scheme, information on the URLLC resource is updated frequently using the control channel of a scheduled user. For example, if an MBB packet consists of 14 symbols, then 10 symbols are used for the eMBB transmission and the rest are reserved for URLLC the Disadvantage of this that when there is no URLLC transmission in the scheduled period the reserved resources for the URLLC will be a wasted. When compared to the semi or half-static reservation scheme, the dynamic reservation requires an additional control overhead to indicate the reservation information. Also, to ensure the reliability of the control signaling, an overhead itself is basically indispensable [29]

II-6 Evaluation Metrics for URLLC Requirements

II-6-1 User Plane Latency

User plane latency (L) is defined as the needed time to deliver an application layer packet from the radio protocol layer entry point until the radio protocol layer exit point in both UL and DL by the radio interface directions, where neither device nor base station reception is restricted by Discontinuous Reception (DRX, a mode in which the UE sleeps for certain periods)

Another definition, user plane latency is the contribution of the radio network to the time from when the source sends a packet to when the destination receives it in ms). Another definition, as the one-way time it takes to deliver an application layer packet/message from the radio protocol layer 2/3 SDU ingress point to the radio protocol layer 2/3 SDU egress point of the radio interface in either downlink/uplink.

The evaluation of 3GPP 5G NR user plane latency is based on taking into account the case of retransmission. If one assumes the latency for initial transmission is To, the latency for initial transmission and onetime retransmission is T, and the latency for initial transmission plus n-time retransmission is T, the expected user plane latency is given by conditions, assuming the mobile station is in the active state.

Similar evaluation method can be applied for URLLC. However, one difference is that for URLLC, usually 99.999% of successful transmission ratio is required. If the air interface design guarantees the initial transmission achieves 99.999% successful ratio, then one can assume that there is no retransmission. In this case, the average delay can be approximated by the delay of initial transmission. [22]

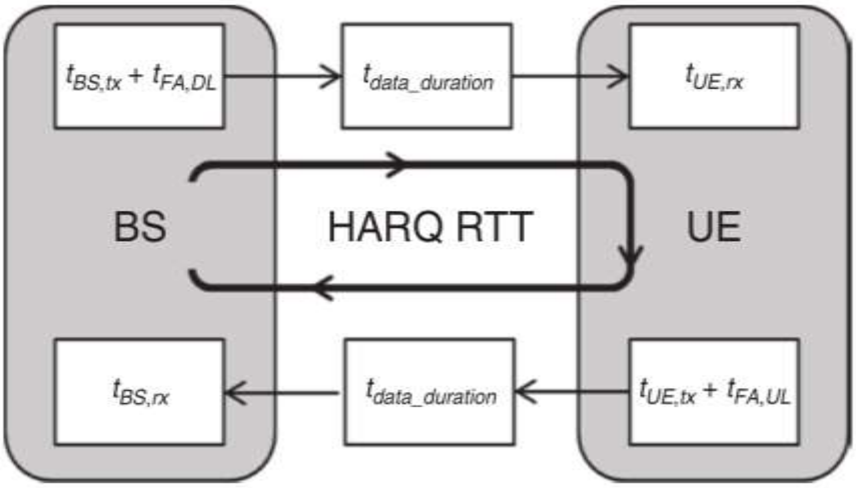


Figure II-4: User plane procedure for evaluation.

$$T(1) = P_0T_0 + P_1T_1 + \cdots + P_NT_N \quad \text{(II-3)}$$

where Pn is the probability of n-times retransmission (n = 0 means initial transmission only), and $P_0 + P_2 + \ldots + Pn = 1$. (II-4)

The probability of retransmission is related to SINR, coding scheme, modulation order, etc. It is noted herein that the exact value of T(l) is dependent on l, which is the index of the OFDM symbol in one slot when the data packet arrives. This is obviously true for TDD band since, if the DL data packet arrives in an UL slot, it of course needs more waiting time for the next DL slot than the case of arriving in the DL slot. For NR, this dependency is also valid for FDD. This is because NR allows sub-slot processing, and if the packet arrives in the later part of the slot, it may need to wait until the beginning of the next slot to further proceed, while if the packet arrives in the early part of the slot, it may probably proceed within the slot. To remove the dependency, an averaged value of user plane latency is helpful. This is defined as follows:

$$T_{UP} = \frac{1}{14\times N} \sum_{I=1}^{14\times N} T(I) \qquad \text{(II-5)}$$

where N is the number of slots that constitutes one period of DL/UL pattern, and 14 is the number of OFDM symbols in one slot. For example, for FDD, N = 1, and for TDD pattern "DDDSU," N = 5. T_{UP} is used in the evaluation.

The definitions and calculations are the same as in DL user plane latency evaluation, and an illustration is found in Figure II-6. It can be seen from the above procedure that, for TDD system, there is frame alignment time which is used to wait for the available time resource (time slot) for the desired link direction, e.g., Step 1.2 and 2.2 for DL or UL procedure. This would result in larger delay of the DL and UL user plane data transmission. Especially, if the downlink dominant frame structure is used, the uplink user plane latency will be poor. DL/UL decoupling mechanism will be helpful in this case. [20]

II-6-2 Control Plane Latency

control plane latency refers to the transition time from a most "battery efficient" state (e.g., Idle state) to the start of continuous data transfer (e.g., Active state). For 3GPP 5G NR, control plane latency can be evaluated from RRC_INACTIVE state to RRC_CONNECTED state. Figure II-6 provides the control plane flow employed in evaluation

The detailed assumption of each step as shown in Figure II-6 is provided in table II-2. The evaluation is for UL data transfer.

It should be noted that the waiting time for the available downlink or uplink time resource for the desired link direction should be calculated in Step 2, 4, 6, and 8. It depends on the detailed DL/UL configuration. Again, if the downlink dominant frame structure is used on a TDD band, the control plane latency might be large. DL/UL decoupling mechanism will be helpful in this case.[20]

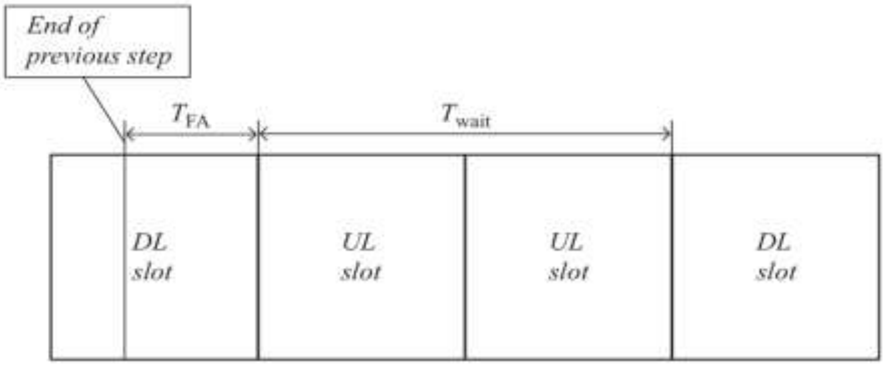


Figure II-5-a: User plane procedure for evaluation.

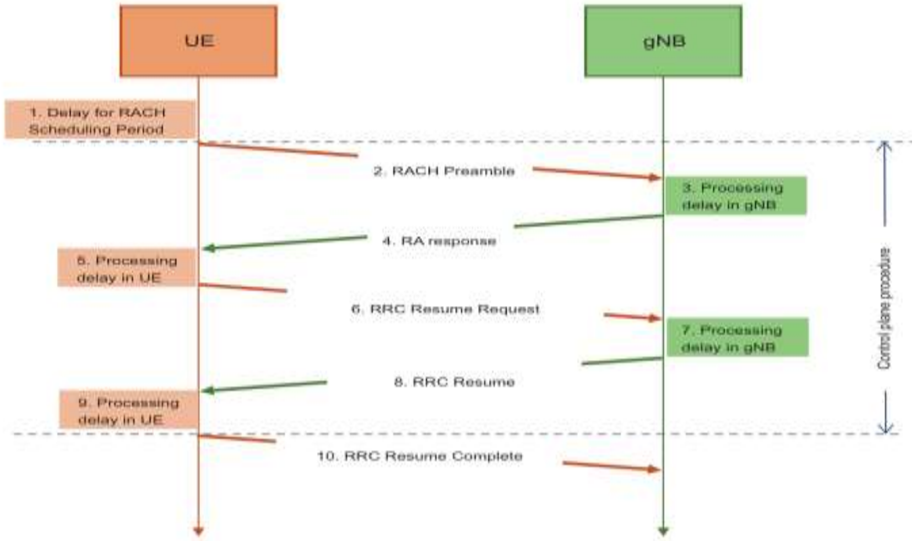


Figure II-5-b: C-plane procedure employed in control plane latency evaluation.

II-6-3 Reliability

Reliability relates to the transmission capability of a traffic within a already determined time duration with based on big success probability.

Reliability is the successful probability transmission in a layer 2/3 packet within a required pre-determined time, which is the time it takes to deliver a small data packet from the radio protocol layer 2/3 SDU ingress point to the radio protocol layer 2/3 SDU egress point of the radio interface at a certain channel quality.[23] This requirement is defined for the purpose of evaluation in the URLLC usage scenario.

The reliability minimal requirement is 1-10−5 success probability of transmitting a layer protocol data unit (PDU) of 32 octets within 1 ms in channel quality of coverage edge for the

Macro-URLLC urban test environment, assuming small application data (e.g. 20 octets application data + protocol overhead). Proponents are encouraged to consider larger packet sizes, e.g. layer 2 PDU size of up to 100 octets.

II-6-4 Mobility

Mobility is the maximum mobile station speed at which a defined QoS can be achieved (in km/h).

The following classes of mobility are defined:

– Stationary: zero km/ hour

– Pedestrian: zero km/ hour to ten km/hour

– Vehicular: ten km/ hour to 120 km/ hour

– High speed vehicular: 120 km/h to five hundred km/ hour

High speed vehicular up to five hundred km/hour is mainly for high speed trains.[22]

II-6-4-1 Mobility Interruption Time

Mobility interruption time is the shortest time duration supported by the system during which a user terminal cannot exchange u-plane packets with any BS during transitions. The mobility interruption time may include the required time to execute any RAN procedure, RRC signaling protocol, or other message exchanges between the UE and the RAN, as applicable to the candidate RIT/SRIT.

This requirement is shown as evaluation in the URLLC and eMBB scenarios. usage the mobility interruption time minimal requirement is 0 ms.

II-6-5 Peak data rate

Peak data rate is the maximum achievable data rate under ideal conditions (in bit/s), which is the received data bits assuming condition of zero errors assignable to a one MS, when all the radio resources for the corresponding link direction are utilized (i.e. excluding reference signals or pilots, radio resources that are used for synchronization of physical layer, guard bands and guard times).

The definition of Peak data rate is for a single MS or UE. In a one band, it is related to the peak spectral efficiency in that band. So Let W_B defined as (channel bandwidth) and S_{PSP} is the peak spectral efficiency in that band. the user peak-data rate R_{peak} is given by:

$$R_{peak} = W_B \times S_{PSP} \quad \text{(II-6)}$$

available bandwidth and Peak spectral efficiency may have different values in various ranges of frequency. In case of where we aggregate the bandwidth across different bands, the peak data rate will be summed over the bands. Therefore, if bandwidth is aggregated across Q bands then the total peak data rate is

$$R = \sum_{i=1}^{Q} W_{B_i} \times S_{PSP_i} \quad \text{(II-7)}$$

where W_Bi and $S_{PSP}i$ (i = 1,…Q) are the component bandwidths , spectral efficiencies respectively.

This requirement is defined for the need of eMBB evaluation in the usage scenario.

The minimum requirements for peak data rate are as follows[20]:

– 20 Gbps for the Downlink peak data-rate

– 10 Gbps for Uplink peak data-rate

II-7 Constraints and Approaches for Achieving Low Latency

In this part we will see the approaches and the constraints of achieving low latency, there are major fundamental trade-offs between capacity, coverage, latency, reliability, and spectral efficiency in a wireless network. Due to these fundamental limits, if one metric is optimized for improvement, this may results in degradation of another metric. Since latency is associated with control overhead (cyclic prefix, transmission mode, and pilot symbols) which occupies a major portion of transmission time of a packet (approximately 0.4 ms per packet transmission), it is not wise to consider a packet with radio transmission time less than 1 ms. If we design a packet with time to transmit of 0.5 ms, more than 60% of the resources will be used by control overhead , Moreover, retransmission per packet transmission takes around 8 ms, and removal of retransmission will affect packet error significantly. As a result, we need radical modifications and enhancements in packet/frame structure and transmission strategy. In this regard: [22]

- First, a novel radio frame reinforced by limited control overhead and smaller transmission time is necessary to be designed. For resource allocation, and channel and channel training can be eliminated or merged. Procedures for user scheduling, reduction of control overhead,

- Second, packet error probability for first transmission has to be minimized with introducing transmission techniques reducing the retransmission delay and new waveforms
- Since latency critical data needs to be dispatched immediately, techniques for priority of data over normal data need to be identified.
- Fourth, synchronization and orthogonality are the indispensable aspects of OFDM that are major barriers for achieving low latency. Even though asynchronous mode of communication is more favorable over synchronized operation in terms of latency, it requires additional spectrum and power resources [22].
- Fifth, since the latency for data transmission also depends on the delay between the core network and the BS, caching networks can be used by storing the popular data at the network edge, to reduce latency.

Researchers proposed various techniques/approaches for achieving low latency in 5G. As summarized in Figure II-7 we divided the existing solutions into three major categories:

1. RAN solutions
2. core network solutions
3. caching solutions.

II-7-1 RAN Solutions

The RAN solutions include new/modified frame or packet structure, waveform designs, multiple access techniques, mmWave modulation and coding schemes, reinforcing Qos and QoE, transmission schemes, control channels enhancements, low latency symbol detection, aggregation, cloud RAN, and location aware communication techniques. energy-aware latency minimization, On the other hand, SDN, NFV, MEC new entities, and fog network along with new backhaul based solutions are proposed for the CN. the caching solution can be subdivided into content delivery, centralized caching, caching placement, and distributed caching, while solutions in backhaul can be divided into mmWave and general backhaul. And in this part we will focus on the RAN solutions provided by researches [22]

II-7-1-1 Frame/packet Structure

In the RAN solutions, modification in the physical air interface has been considered as an attractive choice., most of the solutions proposed are on the medium access control (MAC) layers. And physical In LTE cellular network, the radio frame duration is 10 ms. A frame is divided into 10 subframes of size 1 ms which is further divided into 0.5 ms units that are referred as a resource block (RB). Each RB spans 0.5 ms (6 or 7 OFDM symbols) in time domain and 180 KHz (12 consecutive subcarriers, each of which 15 KHz) in frequency domain. Based on this, the subcarrier spacing (Δf is 15 KHz, the OFDM symbol duration Tofdm is1/Δf = 66.67μs, the FFT size is 2048, the sampling rate fs is

$$\Delta f \times N_{IFFT} = 33.72 \text{ MHz} \quad \text{(II-8)}$$

and the sampling interval T s is 1/f s.

So, to reduce TTI for achieving low latency, the subcarrier spacing Δf can be changed to 30 KHz .This results the corresponding OFDM symbol duration T_{ofdm} to be 33.33 μs and the FFT size N_{fft} to become 1024 while sampling rate fs is kept 30.72 MHz similar to LTE systems. The frame duration T s =10 ms can be divided into 40 subframes in which each subframe duration T sf is 0.25 ms and contains 6 or 7 symbols. Two types of cyclix prefixs (CPs) can be employed in this configuration with durations .

$$T_{cp1} = \frac{5}{64} \times N_{IFFT} \times T_s \approx 2.604\ \mu s \quad \text{(II-9)}$$

$$T_{cp2} = \frac{4}{64} \times N_{IFFT} \times T_s \approx 2.083\ \mu s \quad \text{(II-10)}$$

II-7-1-2 Modulation and Channel Coding

Although use of small packets is a potential approach for achieving low latency, appropriate modulation and coding is required for small packet transmission for acceptable reliability. In the literature, mainly three types of coding schemes are proposed for 5G. As presented in , the (LDPC) or the low-density parity-check and polar codes have a better performance compared to turbo codes in the small packets, but large or medium packets, the opposite is true. While for low latency applications small packet is a requirement, other aspects such as implementation performance in practical test, complexity, and flexibility need to be researched. In polar code has been tested in field for 5G considering various scenarios: air interface, frame, structure, settings for large and small packets, OFDM, and filtered OFDM (f-OFDM) waveforms. In all cases, polar code performed better than turbo codes which makes it a candidate channel coding scheme for 5G. The comparison among the schemes are illustrated in Table II-2 In [23], a highly-parallel architecture for the latency sensitive turbo decoding is proposed by combining two parallel algorithms: the traditional sliding window algorithm and cross parallel window (CPW) algorithm. [26]

Table II-2: comparison among channel coding schemes for low latency.

cases	Turbo coding	LDPC-PEG	CONVOLUTIONAL CODING	POLAR CODES
Algorithm complexity for coding 1/3 of 40bits with respect to turbo codes	100%	98%	66.7%	1.5%
Algorithm complexity for coding 1/3 of 200 bits with respect to turbo codes	100%	98%	66.7%	110.7%
Performance in short packets		✖		✖
Performance in medium packets	✖		✖	

II-7-1-3 Transmitter Adaptation

a system architecture of millimeter-Wave based switching is introduced where signals of control use a low resolution digital beamforming (in order to enable multiplexing of small packets control) with analog beamforming in the data plane (to enable higher order modulation). This reduces the overhead significantly due to the control signaling which results in more resources for data transmission. Recent advancements in full duplex (FD) communication comes forward with advantage of multiplying the capacity by two, also latency mechanism and proving the feedback, in same time upholding steady security od physical layer .Many suggested 5G networks techniques such as beamforming and mMIMO technology, providing reduced spatial domain interference can be helpful for realizing a reliable FD system .also that intelligent scheduling of delay and throughput critical packets along with proper adaption of the rate and power assignment can help to reach a good results in gain of capacity and reduction of latency,however the full duplex Improves throughput, reduces latency and up-holds PHY layer security., but still got some problems with Crosstalk between the transmitter (Tx) and the receiver (Rx), internal interference, fading, and path loss. [27]

II-7-1-4 C-RAN and Other Aspects

Cloud radio access network (CRAN) (as illustrated in Figure II-7**)** is introduced for 5G. CRANs combine baseband processing units of a group of base stations into a central server retaining radio front end at the cell sides. However, this requires connection links with delay of 250 μs to support 5G low latency services. In order to meet strict latency requirements in CRAN, two optimizations, the experimental results clearly demonstrate the effectiveness of the approaches for latency optimization. split of PHY and MAC layer in a CRAN with Ethernet front haul is proposed. The promising results affirm that latency critical services in 5G can be supported by CRAN. Though mmWave will be the major contributor in attaining 5G goals, spectrum below 6 GHz is always the primary choice due to less attenuation, supporting long distance, and antenna compatibility. Moreover, conventional cellular networks are usually deployed within the expensive licensed bands and they use reliable core networks that are optimized to provide low-volume delay-sensitive services such as voice [22] .

a QoE driven dynamic and intelligent spectrum assignment scheme is proposed which can support both cell and device level spectrum allocation. This technique enhances not only the spectrum utilization, but also can maintain desired QoE including latency aspect.

II-7-2 Core Network Solutions

To meet the vision of 5G encompassing ultra-low latency in addition to enhancements in the RAN, drastic changes are also proposed in the core network. The new core network includes some new entities such as SDN, MEC, and NFV as well as new backhaul techniques , These enhancements aim to reduce the processing time, bypass several protocol layers, and ensure seamless operation. The core network solutions for low latency are reviewed in further detail in the rest of this section.

II-7-2-1 Core Network Entities

The SDN and NFV are assumed to be the main candidates for the design of 5G core network . Based on this,we mainly focus on the role of SDN and NFV technologies in latency reduction in 5G core network. The existing literature on the core network entities that can facilitate to achieve low latency are summarized in Table II-3. Decoupling of control plane and data plane seems necessary because they have different network QoS criteria to be met. In particular, the control plane needs low latency to process signaling messages, while the data plane requires high throughput to process the data. Thus, in order to design such planes efficiently, it is preferable to decouple them completely. SDN and NFV can be employed in EPC architecture in order to decouple data plane and control plane and have a seamless operation of core network functions. [22]

Control plane and user plane can be separated by employing SDN in EPC. An SDN controller can act as an interface between the decoupled planes. In addition to several advantages of SDN/NFV-based user plane and control plane separation, including independent scalability, flexibility of flow distribution, and better user mobility management, such a decoupling can have considerable effect on reducing the latency as well. This plane decoupling can facilitate the mobile edge computing technology which decreases the latency. However, adding an SDN controller to the network can be another source of the latency for the system.

Authors proposed SDN/NFV-based MEC networks algorithms that can enable the data plane to create a distributed MEC by placement of network functions at a distributed manner. They demonstrated that the proposed scheme can reduce the redundant data center capacity around 75%, and meet the 5G latency requirement along with considerable backhaul link bandwidth reduction.

Mobility management in core network based on SDN can potentially introduce some delays. In the main contributors for processing delays in an SDN-based mobility management system is discussed. By implementing two proactive and reactive solutions for mobility management using Mininet and Open Flow, it is observed that with high probability (almost 95%) in the proactive mobility management system, the overall processing latency is around the median value.

II-7-2-2 Backhaul

Backhaul between base stations and the core network carries the signaling and data from the core and the Internet. Due to the enormous number of small cells and macro cell base stations supporting 1000x capacity, massive connectivity and latency critical services in 5G, the capacity of backhaul is a bottleneck for achieving low latency. At current scenario, microwave, copper and optical fiber links are used for backhaul connections based on availability and requirements. 5G backhaul requires higher capacity, lower latency, synchronization, security, and resiliency

II-7-2-3 Caching Solutions

In addition to the shortage of the radio spectrum, the insufficient capacity of backhaul links can be considered as a bottleneck for low latency communication. The long delay can be due to the requests of too many users in peak-traffic hours. Thus, latency reduction is crucial for users' QoS and QoE in the 5G networks. Caching and in a more general category, information centric networking, can be assumed as one of the promising candidate technologies to design a paradigm shift for latency reduction in next generation communication systems we present a detailed overview of caching concepts for cellular network followed by fundamental limits and existing solutions.[22]

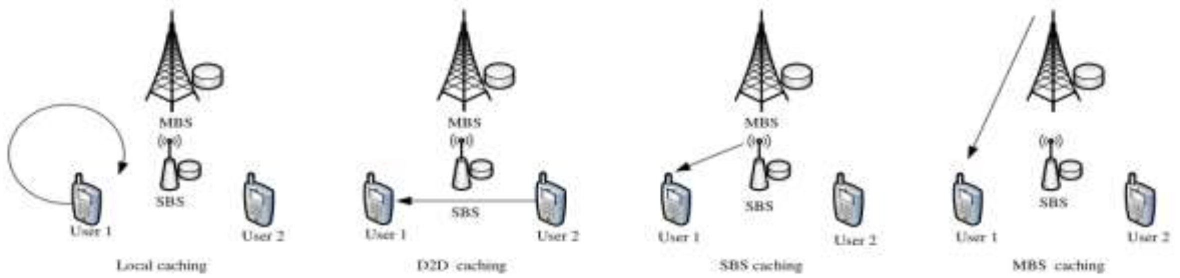


Figure II-6: Different types of caching in 5G.

In summary, A different architectures are suggested for designing the core network of the 5G, the Table II-3 illustrates them:

Table II-3: overview of techniques in core network for low latency.

Case/area	Approach	Summary
CORE NETWORK ARCHITECTURE	SDN-BASED	The architecture of 5G is proposed based on SDN with 5G vision to meet large throughput, massive connectivity and low latency
	NFV-BASED	NFV invalidates the dependency on hardware platform and makes easy deployment of PEC functions as well as the sharing of resources in the RAN, this can reduce E2E latency with improved throughput performance
	MEC/FOG based	MEC/fog provides computation and storage near user end and also separates the data plan from control plan, these reduces the latency

In general, we can gather all the suggested solutions for a low latency in this figure:

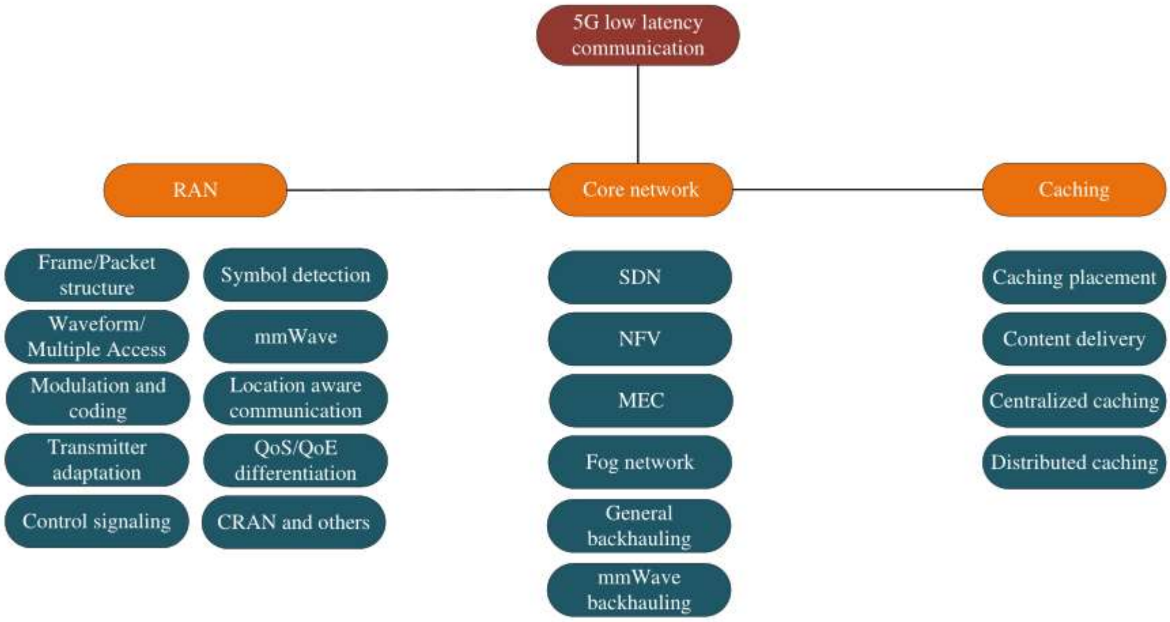


Figure II-7: Categories of different solutions for achieving low latency in 5G.

II-8 Conclusion

In addition to extremely wide bandwidth, latency is also related to it ,as the response time of a system ,it could be combined with speed and ultra-high coverage, 5G networks would need to accommodate ultra-low latency. Low latency would allow digital technologies such as VR / AR, telemedicine and telesurgery; in certain situations, latency is not more than 1 ms essential. Numerous improvements in various network contexts need to be tackled in order to maintain this low latency. A detailed study of various solutions for achieving low latency in 5G networks is also provided with specific strategies in the area of RAN, core network and caching for achieving low latency. In the area of RAN technologies, we have researched short frames / packets, modern waveform architectures, multiple access strategies, modulation and coding systems , control channel approaches, symbol recognition methods, propagation technologies, mmWave amplification, cloud RAN, QoS-enhancing and location-conscious connectivity as different aspects of promoting low latency.

Despite its significance, physical layer technologies are struggling to incorporate URLLC into SG NR., In order to fulfill the strict latency and durability criteria, certain aspects of the physical layer need to be revamped, there are other fascinating things worth investigating, such as the beam-forming control and data component technique and the reconfigurable portion. URLLC protocol. Also, study of advanced transceiver architecture to support dynamic numerology adaptation and simultaneous decoding is needed.

At the other side, SDN, NFV network models coupled with high-speed backhaul was analyzed in the literature for the core network with aim to satisfy the low latency criteria of 5G. The latest core network can deliver a range of benefits, such as distributed network design , software device independence from hardware framework, and the isolation of the data plane from the device plane, both of which would help to reduce latency. Throughout this regard, we addressed open topics, problems and potential directions for work for researchers.

Chapter III: Frame Structure Enables Latency in 5G

III-1 Introduction

As we have seen in chapter 2 ,In order to achieve latency ,many aspects need optimization and modification ,Thus , this chapter mainly focus about the Radio Area Network ,including physical layer technologies ,air interface technologies , as an overview , This gives a review for paramount waveform outline prerequisites to diverse situations Mobile broadband has high performance specifications and low latency, generally small cells at higher carrier frequencies (mmwave), larger cells at lower carrier frequencies.. For the downlink, high spectral quality and MIMO usage capacity are main criteria for the NR waveform. Nevertheless, in the case of uplinks in large cell deployments, the cell maximum or edge of the UEs can be constrained in power. For this scenario, the power efficiency of the The waveform is quite significant. For small cells, the distance between the UEs and the BS may be reduced dramatically.

Carrier frequencies mm-wave, analog or hybrid beam forming are required. For RF beam forming, it is not feasible to use multiplex UEs (Frequency Division Multiplexing (FDM)) frequency domain, unless the same beam serves all of them. In this case, the users use (time division multiplexing (TDM)) so, they are all multiplexed in the time domain and therefore a short transmit time interval (TTI) is Critical. The NR waveform short transmission has to be confined in time.

URLLC is necessary to increase reliability and to obtain very low latency, the waveform should be low latency processing and confined to the time domain. Asynchronous communication would be useful for quicker uplink connectivity, essentially the waveform would have the capacity to accommodate asynchronous connectivity. Enhancing the stability of the link by implementing the retransmission mechanisms.

The key performance indicators are used in drive test for evaluating a radio system , and also ,we can use them as a tool of evaluation in the physical layer , since latency is not one thing , we need to perform several tests on many corners of the RAN ,as well as the physical layer in general .

III-2 Key Performance Indicator for 5G NR Waveform Design

The characteristics of 5G NR including extreme data rates large channel bandwidths, harsh propagation conditions, URLLC requirements, severe RF impairments, small sized base stations, antenna's number is massive, and for TDD deployments. The 3GPP identified key performance indicators (KPIs) for the waveform design ,include [30] :

• **Spectral efficiency:** The spectral efficiency is the number one important to support requirements on traffic densities, data rate, and user connection. In general, lower carrier frequencies has much more important to the KPI of spectral efficiency than at higher frequencies

• **MIMO compatibility:** because of the increment in carrier frequency, the antenna elements number would increase in the BS as well as in the UEs. The use of different schemes for MIMO is important in giving high spectral efficiencies and glad to beamforming we assure greater coverage .

• **Low PAPR (Peak-to-Average-Power-Ratio):** the power efficient transmissions is important as a low PAPR is essential for it, from the devices especially for the uplink side (UEs). At very high frequencies a low PAPR becomes even more essential. Since small sized low-cost BSs are envisioned at high frequencies, but also, low PAPR is also could be important for the down link transmissions.

• **Robustness to channel time selectivity:** The high speed scenarios are shown in large cells. Due to the wave obstacles for propagation conditions, The large cells might not be a good choice at very high frequencies, where it might cause a limitation in the coverage. The deployments are expected to come in the form of small cells, at very high frequencies, where mobility is not a major concern.

• **Transceiver baseband complexity:** The baseband complexity is always very important, especially from the receiver point of view. For NR, the complexity is even a major consideration for BSs, since at high frequencies a BS can be a small sized access node with limited processing capability , for faster processing and enabling low latency applications A low baseband complexity is also important.

• **Time localization:** in dynamic TDD mode ,the localization in time is crucial for making the mode supports a low latency applications. The TDD's frequent link direction switching short burst transmissions is one of the requirements . Also this needs a short transmission time-slots which helps enabling a low latency, it's one of the key requirements for both the eMBB and the URLLC,. for enabling short transmissions a waveform that is confined in the time domain is suitable .

• **Out-of-band emissions and Frequency localization:** in FDD mode, the localization in Frequency is useful for potential multiplexing and efficient utilization of spectrum in different 5G NR applications, using various waveform numerologies at high frequencies, Frequency localization is not a major KPI where large channel bandwidths are available. Frequency localization can be helpful in both uplink and side link which is also relevant for asynchronous access

• **Scalability and Flexibility** Scalability and flexibility of the waveform is needed for much more improvement in the future

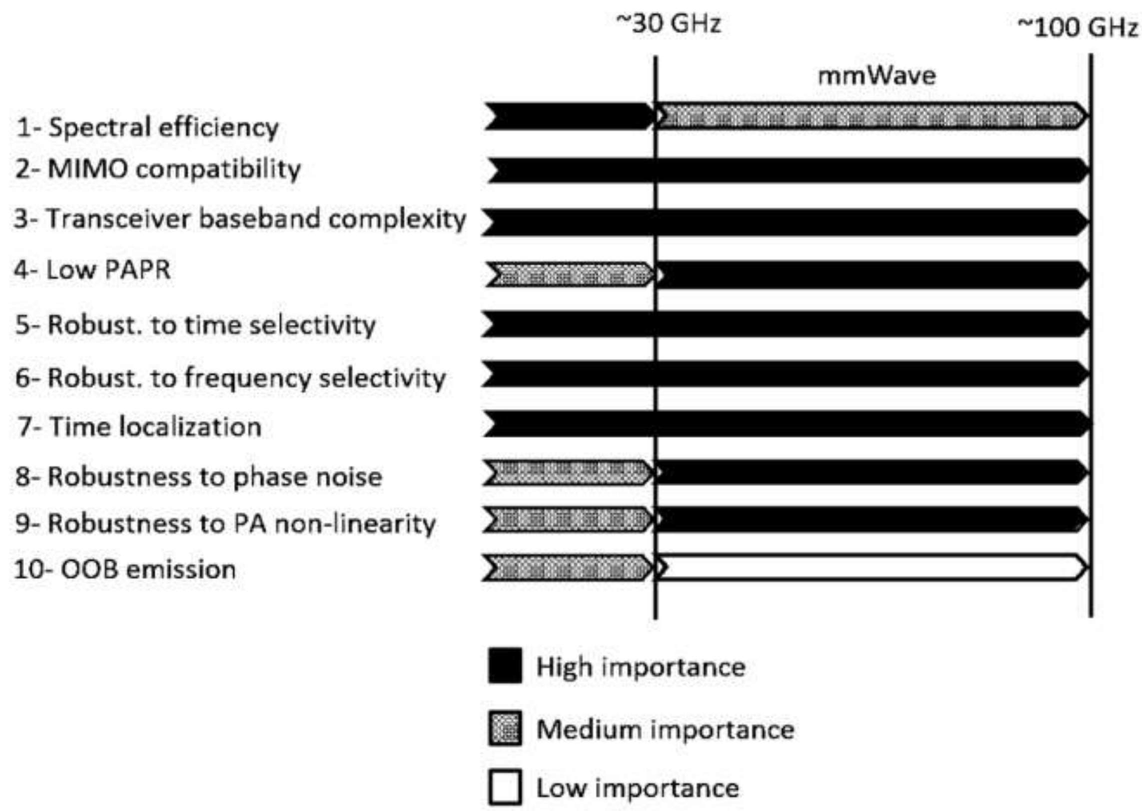


Figure III-1: Importance of waveform performance indicators especially for low latency.

Figure III-1 shows the importance of the waveform Key Performance Indicators for the NR in different frequency ranges. We can see that for mmwave communication, special attention needs to be paid to much more hardware impairments and power efficiency, whereas frequency localization is not of great importance

III-3 Waveform Comparison For 5G NR

Table III-1: Comparison summary of multicarrier waveforms.

MC-waveforms	CP-OFDM	W-OFDM	UF-OFDM	FBMC-QAM	FBMC-OQAM
Spectral efficiency	High	High	High	High	High
PAPR	High	High	High	High	High
Robust. to time-selective channel	Medium	Medium	open	Low/ Medium	Low/ Medium
MIMO compatibility	High	High	Open	open	Low
Time localization	High	High	High	Low	Low
OOB emissions	High	Medium	Medium	Low	Low
OOB emissions with PA	High	High/ Medium	High/ Medium	High/ Medium	High/ Medium
Complexity	Low	Low	Medium	High	High
Flexibility	High	High	High	High	High

In this comparison , we evaluated different waveform candidates, multicarrier and single-carrier, The waveform comparison results are summarized in Table III-1 for selected multicarrier waveforms. The following color scheme has been used in this table: **Green** (light gray in print version) refers to a desirable characteristic; **Red** (medium gray in print version) refers to an undesirable characteristic; **Blue** (dark gray in print version) refers to somewhere between desirable and undesirable characteristic. Moreover, "Open", means that further investigations are required to draw any conclusion. The investigations were based on analytical methods as well as simulations. The simulations were performed under common evaluation assumptions in the mmMAGIC waveform simulators, The detailed evaluation results are available in the mmMAGIC project report, we provide the waveform comparison results for selected KPIs (power efficiency, phasenoise robustness, time localization, baseband complexity) and refer the reader to [30] for further details.

III-3-1 Power Efficiency

a common drawback of all multicarrier waveforms is their high PAPR (and low power efficiency). In Figure III-2, we compare the PAPR of several waveforms, including CP-OFDM,

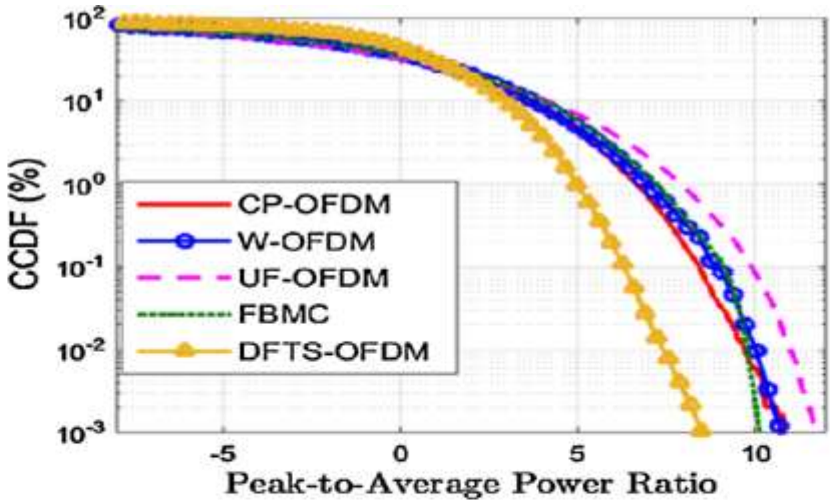


Figure III-2: A comparison of PAPR of multicarrier waveforms and single carrier DFTS-OFDM waveform.

W-OFDM, UF-OFDM, FBMC-OQAM, and DFTS-OFDM (assuming 16 QAM and 1200 subcarriers), We observe that all multicarrier waveforms have similar PAPR except UF-OFDM, which has a higher PAPR. The DFT-based precoding in OFDM (DFTS-OFDM) reduces the PAPR and achieves a higher power efficiency than OFDM. There are various well-known methods to improve the power efficiency of OFDM [30]

III-3-2 Time-Varying Fading Channel

We compare performance of several multicarrier waveforms (CP-OFDM, W-OFDM, UF. OFDM, FBMC-OQAM, and FBMC-QAM) as shown in Figure III-3, in terms of the symbol error rate over a time-varying fading channel with 60 km/h UE speed at 6 GHz carrier frequency (assuming QuaDRiGa channel model). For all waveforms, assuming 120 MHz Signal bandwidth, 16-QAM ,and 512 subcarriers. As can be seen. CP-OFDM has better performance compared to other waveforms in the time varying channel. fading

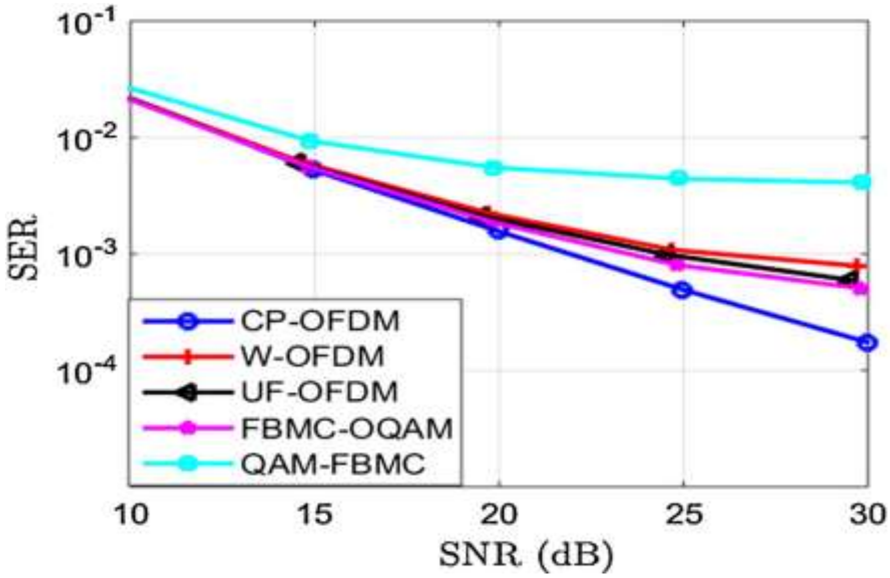


Figure III-3: A comparison of multicarrier waveforms subject to a time-varying fading channel (60 km/h UE speed at 6 GHz carrier frequency).

III-3-3 Baseband Complexity

The multicarrier waveforms implementation complexity in terms of the number of real multiplications required for demodulation and synthesis, without the computations required for channel estimation. There are many techniques for channel estimation used in practice with varying degrees of complexity, We compare complexity of the following multicarrier waveforms: OFDM. W OFDM F-OFDM, UF-OFDM and FBMC. Since FFT / IFFT operations are much more used in implementations of all of these suggested waveforms, for much more simplicity we will denote the complexity of an N-point FFT/IFFT operation with C_{FFT} (N).

III-4 Suitability of OFDM for 5GNR

For the NR waveform, the design requirements vary depending on the carrier-frequency range and the link type. After The 3GPP investigation of many multicarrier/single-carrier waveforms, the candidature CP-OFDM waveform has been selected for NR after considering the requirement criteria for the NR design, also, it provided a comparison of state-of-the-art waveforms. so we will discuss the CP-OFDM and evaluate it based on the KPIs [30] :

• **Spectral efficiency:** spectral efficiency is known in the OFDM for its high. The spectral efficiency is essential to meet the needed requirement for data rate with NR. In general, as we said before for lower carrier frequencies the spectral efficiency is more crucial rather than at higher frequencies due to potentially much larger channel bandwidths at higher carrier frequencies. The spectral efficiency plays a big role for both uplink/downlink. Because of the huge amount of data transferred between the BSs, VANET can also create capacity in such scenarios as dense urban area with numerous vehicles periodically broadcasting signals in an asynchronous fashion.

•**MIMO compatibility:** technology to increase in the carrier frequency with NR , OFDM enables a straightforward use of MIMO, the number of antenna elements would increase in the BSs as well as in the devices. by enabling SU-MIMO/MU-MIMO The use of various MIMO schemes is essential for enhancing spectral efficiency and achieving greater coverage via beam-forming,

•**Peak-to-Average-Power-Ratio (PAPR) :** the drawback of OFDM that it has a high PAPR compared to other multicarrier waveforms ,as shown in Figure III-2, A low PAPR is essential for power efficient transmissions from the UEs uplink. A low PAPR becomes even more important at very high frequencies where coverage can be limited. low PAPR is also important for the downlink transmissions at high carrier frequencies because It is noticeable that small sized low cost BS are envisioned at high frequencies

For NR, OFDM with PAPR reduction is an attractive option for the uplink and the sidelink. LTE uses DFTS-OFDM for both uplink and sidelink due to its lower PAPR. However, DFTS-OFDM has certain drawbacks in comparison with OFDM such as lesser flexibility for scheduling and more complex MIMO receiver with degraded link level and system level performance .

For NR uplink the DFTS-OFDM is optional, and can be used only for a single-stream (without MIMO) transmission. DFTS-OFDM is not a preferred option for the uplink Since MIMO transmission is a key component of NR,.

•**Robustness against channel time selectivity:** OFDM can be made robust to channel time selectivity by choosing of right subcarrier spacing depends on the condition and application ., furthermore , for the envisioned mobile backhaul (ex : the access nodes) on vehicles, the robustness is much more relevant .

in high mobility scenarios the High-speed users are relevant in large cells, due to harsh propagation conditions The large cell deployments are not widely employed at very high frequencies because of coverage limitation. However, the vehicle to vehicle could be applied at very high frequencies, making the selectivity in the channel time robust where the mobility is not big concern in this case because its basically fixed,

•**Time localization OFDM:** The localization in the time efficiently helps to apply the TDD mode for supporting the low latency applications, and provision of it is essential for all link types, especially V2V and backhaul links need a higher requirements.

•**Robustness to the synchronization errors:** The existence of the cyclic prefix (CP) makes OFDM much more robust to errors in timing synchronization Robustness to the synchronization errors, especially in cases where the synchronization is hard to achieve such as in the sidelink. It can also be important if asynchronous transmissions are allowed in the uplink. (where LTE only supports synchronous uplink transmission, which is realized via a timing advance at the user end)

• **Flexibility and scalability:** a proper choice of the subcarrier spacing and the cyclic prefix. Is what makes OFDM as a flexible waveform which can support diverse services in a wide range of frequencies ,which what we are going to simulate down below ,and show how the waveform enables low latency by the flexibility of the waveform design

in Figure III-4. a summary of different link types for the waveform design requirements is shown , A link requirement "High" for a waveform KPI tells that the given KPI is important for the given link type in general. Furthermore, a high-level assessment of OFDM is given in Table III-2. The OFDM assessment "High" means that OFDM has good performance in general for the given KPI. We assess the D2D and V2V cases separately due to different levels of requirements. As an example, V2V communication in term of the mobility ,has a higher requirement , and the system capacity, whereas for the power efficiency ,there are lower requirements , when V2V is compared with UE-to-UE communication . Based on the overall assessment, it can be concluded that OFDM is an excellent choice for the NR air interface for all link types.

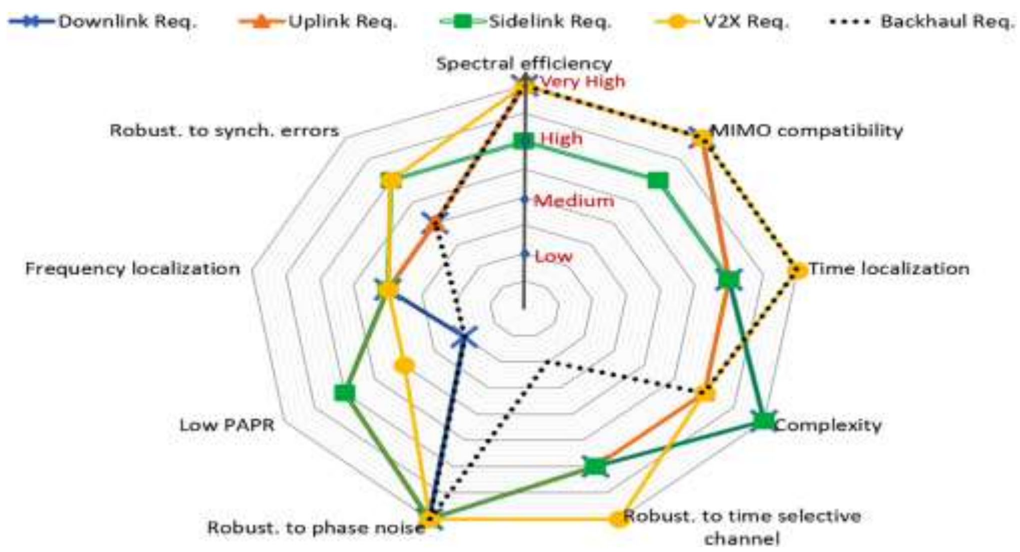


Figure III-4: A high-level summary of the waveform design requirements for different link types: uplink, downlink, sidelink,,V2V link, and backhaul link.

Table III-2: A high-level assessment of OFDM on latency.

Performance indicators	OFDM assessment
Spectral efficiency	**High**
MIMO compatibility	**High**
Time localization	**High**
Transceiver baseband complexity	**low**
Flexibility/Scalability	**High**
Robust. to time-selective chan	**medium**
Robust. to synch. errors	**medium**
PAPR	**High** (can be improved)

III-5 NOMA vs OMA

After comparing the Non orthogonal wavesforms,now ,we would try to illustrate major differences between non orthogonal multiple access (NOMA) and orthogonal multiple access (OMA)

III-5-1 NOMA

Multiple access lies at the heart of cellular communication systems. It refers to a technique that allows multiple users to share a communication channel. The first- generation (1G) to the fourth-generation (4G) of cellular networks have adopted radically different multiple access schemes with one common theme in mind—to have orthogonal signals for different users at the receiver side Table III-2 in [30]. In particular, 1G to 4G cellular networks has adopted one or more of the following multiple access

methods:

- Frequency division multiple access (FDMA)
- Time division multiple access (TDMA)
- Code division multiple access (CDMA)
- Orthogonal frequency division multiple access (OFDMA)
- Space division multiple access (SDMA)

For example, in OFDMA which has widely been used in 4G systems, different users' signals are orthogonal in the frequency and/or time domains. In other words, one orthogonal frequency division multiplexing (OFDM) resource block (180 kHz) cannot be allocated to more that one user.

Non-orthogonal multiple access (NOMA) [31] ,in contrast, allows multiple users to share the same resource elements, be it in the time, frequency, space, or code domain. NOMA is currently a hot research topic for 5G and beyond systems, both in academia and industry. While it is concerned with "non-orthogonality" of multiple access, it appears that the research community is perceiving this term in somewhat different ways. Due to the different interpretations, there is not a consensus about applying this term to some well-known existing techniques such a CDMA. While the majority of recent works see CDMA as an orthogonal multiple access (OMA) technique, there are a group of other papers that categorize it as a NOMA technique. In the following, we present and discuss different viewpoints used to define non-orthogonality in NOMA.

III-5-2 NOMA Two Users Scenario

Consider two users example [32]:

The base station selects two appropriate users to pair, as shown in the figure

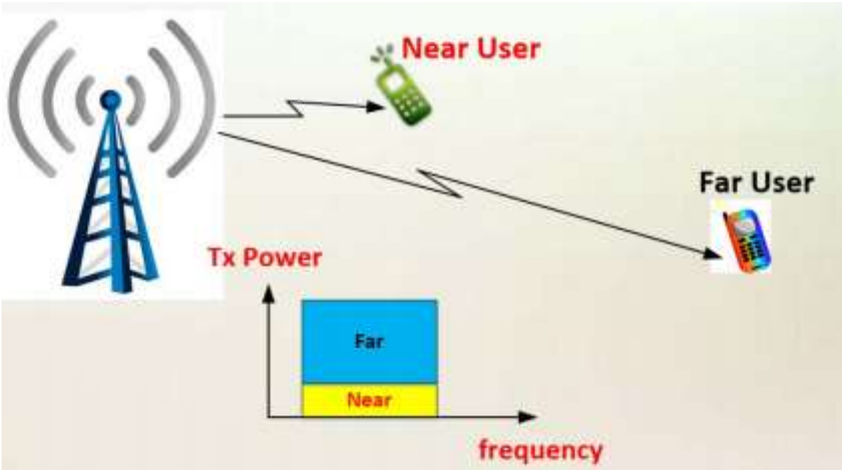


Figure III-5-a: two users example using NOMA.

Where the Near user has (strong channel gain) and the Far user has (weak channel gain) Where the modulation scheme of NOMA using QPSK(QAM-4) is different than usual modulation QPSK as shown in Figure III-5-b

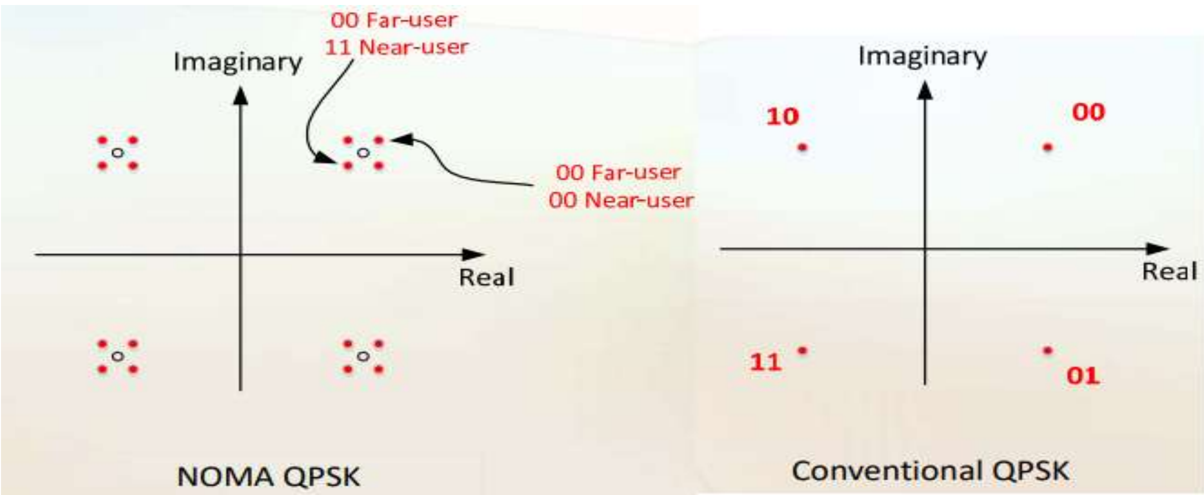


Figure III-5-b: NOMA QPSK constellation diagram.

We note that :

- Served on same time and frequency
- TX Power is split between them
- High power share to Far-user
- Low power share to Near-user

So Typically interference should happen !

III-5-3 Mechanism of NOMA

Far-user signal has small interference from the Near-user signal, and the Far-user decodes its signal normally, as shown in the figure Figure III-7-a. But it Suffers from slight extra interference [32] :

1- Near-user signal has large interference from Far-user , Near user decodes Far-user signal first, as shown in figure Figure III-7-b

2- Subtracts this interference from the composite NOMA signal

3- Far-user interference is cancelled

4- Near-user decodes its data from the cleaned signal

Note: the Far-user is unable to cancel Near-user interference because it is too weak to be decoded.

Figure III-7-a: Far-user decoding scheme

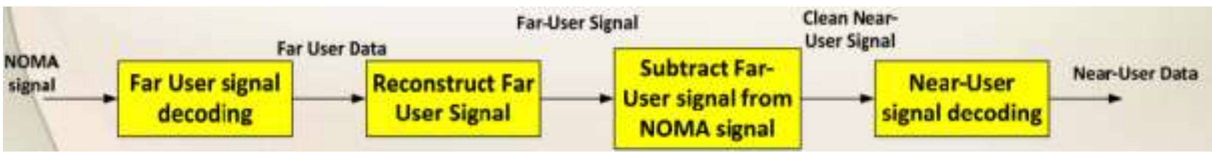


Figure III-7-b: Near user decoding scheme

Capacity of NOMA is the major advantage over all the advantages of OMA techniques ,using Shannon law for capacity to evaluate NOMA capacity

$$C = w \times \log_2(1 + SNR) \quad \text{bit / s} \qquad \text{(III-1)}$$

Where The bandwidth (w) has the highest impact on the capacity .where in NOMA ,the bandwidth is divided by the near user and the far user ,as shown in the figure 8 ,otherwise in NOMA the bandwidth is still the same shared to the two users

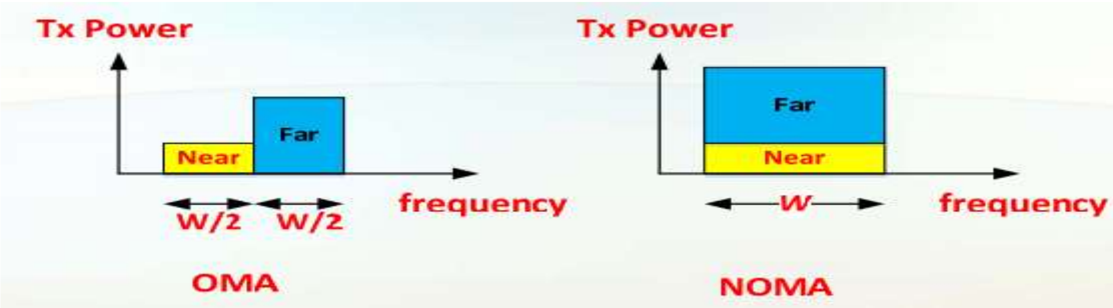


Figure III-8: OMA vs NOMA in bandwidth allocation.

In OMA ,each user use half the bandwidth ,but without any interference which assures reliability for this technique[32] :

In OMA :

$$OMA_{SNR} = \frac{P_{near}}{noise} \qquad \text{(III-2)}$$

And $$SNR_{FAR} = \frac{P_{FAR}}{noise} \ . \qquad \text{(III-3)}$$

We can see that ,the NOMA near-user can cancel far-user signal :

$$SNR_{near} = \frac{P_{near}}{noise} \qquad \text{(III-4)}$$

but NOMA far-user can't cancel Near-user signal,because we have additional noise :

$$SNR_{FAR} = \frac{P_{FAR}}{P_{NEAR} + noise} \ . \qquad \text{(III-5)}$$

So , the capacity of both NOMA and OMA would be :

$$C_{OMA} = \frac{W}{2} LOG_2\left(1 + \frac{P_{near}}{noise}\right) + \frac{W}{2} LOG_2\left(1 + \frac{P_{far}}{noise}\right) bits/s \qquad \text{(III-6)}$$

$$C_{NOMA} = W \ LOG_2\left(1 + \frac{P_{near}}{noise}\right) + W \ LOG_2\left(1 + \frac{P_{far}}{P_{near} + noise}\right) bits/s \qquad \text{(III-7)}$$

Where : $$P_{far} + P_{near} = P_{total} = constant \qquad \text{(III-8)}$$

III-5-4 NOMA viewpoints

Major advantages of NOMA compared to OMA Serve more than one user on the same time and frequency resource

• Higher spectral efficiency (more data rate per Hz)

• Benefit from the geographical distribution of users

• Better serve cell edge users (users far from the base station)

• Some applications need low data rate. A waste of resources to

allocate dedicated time and frequency

• Sensor readings

• Inter-vehicle communications

• Machine to machine communications

• **Superposition Coding with Successive Interference Cancellation**: A large body of papers consider NOMA to be equivalent to superposition coding and successive interference cancellation (SC-SIC), respectively, at the transmitter and receivers. This is partly because the first paper using the term NOMA considered the problem of downlink transmission with SIC and partly due to the fact that SC-SIC is the capacity-achieving technique for the downlink channel in single-cell single-input single-output (SISO) transmission [31]

• **Overloading:** A second important view is to distinguish NOMA and OMA Based on the system loading. In this setting, NOMA refers to overloaded systems and "overloading" means to have more than one user per available resource element in the time, frequency, code, or space domain. This point of view is rooted in CDMA systems. With this definition, a CDMA system will be seen as a NOMA scheme if it is overloaded (i.e. when there are more users than the number of codes) and will be considered as an OMA scheme if there are more codes than the number of users in the systems. Examples of the NOMA schemes developed with this view are low-density spreading (LDS) CDMA, LDS-OFDM

• **Linear Transform Decoding:** Some even define NOMA based on the complexity of multi-user detection. With this point of view, in an OMA scheme, the signals of different users can be separated in orthogonal subspaces using a linear transform. Then, any scheme that does meet this definition can be categorized as NOMA times for the three operators: approximately

40 ms in the median, and up to 200 ms in the 99%-percentile. The observed performance is also in line with the reported in other field test campaigns

III-5-5 What Drives NOMA

Effective multiple access is a key enabler in achieving 5G requirements, in term of latency and reliability as well as higher throughput , As noted above, the first to the fourth-generations of cellular networks have adopted different multiple access schemes with one common theme in mind—to have orthogonal signals for different users at the receiver side. By allowing multiple users to share all domains (e.g. time, frequency, space), NOMA can address the above challenges of the next generation of wireless networks more efficiently than the conventional orthogonal multiple access schemes. NOMA can increase spectral efficiency and user fairness by exploiting a capacity-achieving scheme in the downlink,. It can support more connections in the uplink by letting multiple users simultaneously access the same wireless resources, which, in turn, can reduce latency for the great capacity, but in term of reliability, OMA still guarantee a user bandwidth.

The following table summarize major differences between NOMA and OMA [33]

Table III-3: NOMA vs OMA Comparison.

Specifications	**NOMA**	**OMA**
Full Form	Non-Orthogonal Multiple Access	Orthogonal Multiple Access
Receiver complexity	High	Low
Energy consumption	More	Less
Number of users/cluster	Lower	Higher
Number of user pairs	Less	More
System throughput	Larger	Smaller

III-6 Frame Structure of 5G NR

Physical layer is the first layer of the NR radio interface protocol architecture. Physical layer provides data transport services to higher layers. Transport services are accessed through transport channels between the physical layer and the medium access control (MAC) part of layer two. The data transport service of the physical layer includes for example coding, modulation/demodulation, multi-antenna processing and mapping of the coded transport channel to the correct physical channels. In addition to the layer two, the physical layer also interfaces with the radio resource control (RRC) part of the layer three.

The NR uses orthogonal frequency-division multiplexing (OFDM) as a waveform in both the uplink and the downlink transmission directions .OFDM provides a flexible multiple-access scheme, where the available bandwidth (BW) is divided into mutually orthogonal subcarriers, which can be shared amongst multiple users without intra-cell interference. OFDM is nowadays a broadly adopted technology that is suitable also for the NR because of the merits like low complexity, plain channel estimation and easy multiple-input and multiple-output (MIMO) -integration . In addition, the uplink transmission scheme has also a possibility to use discrete Fourier transform (DFT) -precoded OFDM as a waveform. DFT-precoded OFDM is used to obtain higher power-amplifier efficiency. However, DFT-precoded OFDM is used only as a complementary support, because of the several drawbacks. DFT-precoding increases the complexity of MIMO-receivers and restricts scheduling in the frequency domain. [34].

With only one OFDM numerology, the NR could not fulfil the performance requirements of the wide range of applications, devices and carrier frequencies For this reason, the NR supports multiple OFDM numerologies with a range of subcarrier spacings (SCSs). The supported subcarrier spacings are 15, 30, 60, 120 and 240 kHz, but 240 kHz is only supported by the synchronization signal block (SSB), which is further described in Section 2.2.2. The flexible numerology is based on scaling, where subcarrier spacing of 15 kHz is the baseline. The baseline of 15 kHz was chosen, so that the NR supports efficient coexistence with LTE .Scaling of the subcarrier spacing from the baseline is defined as

$$\Delta f = 2^{\mu} \times 15 \text{ kHz} \qquad \text{(III-9)}$$

Where :

$$Tu = \frac{1}{SCS} \qquad \text{(III-10)}$$

$$SCS = \Delta f \qquad \text{(III-11)}$$

Where the subcarrier spacing configuration μ is an integer between 0-4 and comes from the higher-layer parameter subcarrier Spacing [35]. Subcarrier spacing with proportional symbol and cyclic prefix (CP) durations are presented in Table III-5,

Table III-4: Supported subcarrier spacing with proportional symbol and CP durations.

Subcarrier spacing configuration (μ)	Δf [kHz]	Symbol duration [μs]	Cyclic prefix duration [μs]
0	15	66.7	4.7
1	30	33.3	2.3
2	60	16.7	1.2
3	120	8.33	0.59
4	240	4.17	0.29

NR transmissions are organized into 10 ms long frames. Each frame is then divided into ten subframe and they are further divided into slots. Each slot consists of 14 OFDM symbols in the time domain and 12 subcarriers in the frequency domain. An exception to this is the 60 kHz subcarrier spacing with an extended CP, where each slot consists of 12 OFDM symbols and 12 subcarriers. The extended CP for 60 kHz can be used when it is needed to reduce the slot duration and delay while maintaining a CP similar to 15 kHz. In the time domain, the slot length in milliseconds is defined as:

$$Slot\ length\ = {}^{1\ ms}/_{2^{\mu}} \tag{III-12}$$

Where μ is the subcarrier spacing configuration. Based on slot length . The number and the length of slots with the supported SCSs is presented in Table III-6 When the subcarrier spacing is doubled, the number of slots within a subframe is also doubled.[34]

Table III-5: Number and length of slots with the supported subcarrier spacing.

Subcarrier spacing configuration (μ)	Number of symbols in a slot	Number of slots in a subframe	Number of slots in a frame	Slot length [ms]
0	14	1	10	1
1	14	2	20	0.5
2	14 (normal CP) 12 (extended CP)	4	40	0.25
3	14	8	80	0.125
4	14	16	160	0.0625

5G NR allows transmissions through different system bandwidths depending on the frequency range. Frequency Range I (FRI) is defined from 450 MHz to 6 GHz and allows system bandwidths of 5, 10, 15, 20, 25, 40, 50, 60, 80 or 100 MHz Frequency Range 2 (FR2) is specified from 24.25 GHz to 52.6 GHz and enables bandwidths of 50, 100, 200 or 400 MHz

The combination of one subcarrier allocated in one OFDM symbol is defined as one Resource Element. A group of 12 consecutive subcarriers in frequency domain is defined as Resource Block. Based on all these parameters, an illustrative example of the framing structure for numerology μ= 0 is shown in Figure III-9.

The time-division multiplexing (TDM) scheme in the NR is more flexible than for example TDM in LTE ,OFDM symbols in a slot are classified as 'uplink'',downlink' or 'flexible. The flexible slot is a combination of uplink and downlink symbols,The UE assumes that the downlink transmission occurs only in downlink' or 'flexible' symbols and thus it transmits only in 'uplink' or 'flexible' symbols .[36]

In addition to the three configured slot types, the NR supports so-called mini-slots. Minislots can be used to support shorter transmissions, since a mini-slot can be as short as one OFDM symbol in time. A mini-slot can also start at any time within a slot to enable a more flexible start position in the time domain. In order to obtain low latency, the beginning of the slot can also be front-loaded with control and reference signals (RSs). Because of the flexible time structure of the mini-slots, they can be used for example in low-latency scenarios, where the transmission needs to start immediately.

In the frequency domain, physical resource blocks (PRBs) are the basic scheduling units of the NR One PRB is composed of 12 subcarriers, which have the same subcarrier spacing and CP overhead within a PRB. Each individual element, grouped into one subcarrier of the PRB, is called a resource element (RE) .The RE is the smallest unit in the resource grid. It corresponds to a physical resource made up of one OFDM symbol in time and one subcarrier in frequency. So, a PRB is composed of 12 REs and they are uniquely identified in the frequency/time domain of the PRB

An example of the NR frame structure with the subcarrier spacing of 15, 30 & 60 kHz is presented in Figure III-9. It illustrates how doubling the SCS doubles the bandwidth and halves the slot duration. As can be seen, with a higher SCS, more slots fit into one subframe. [36]

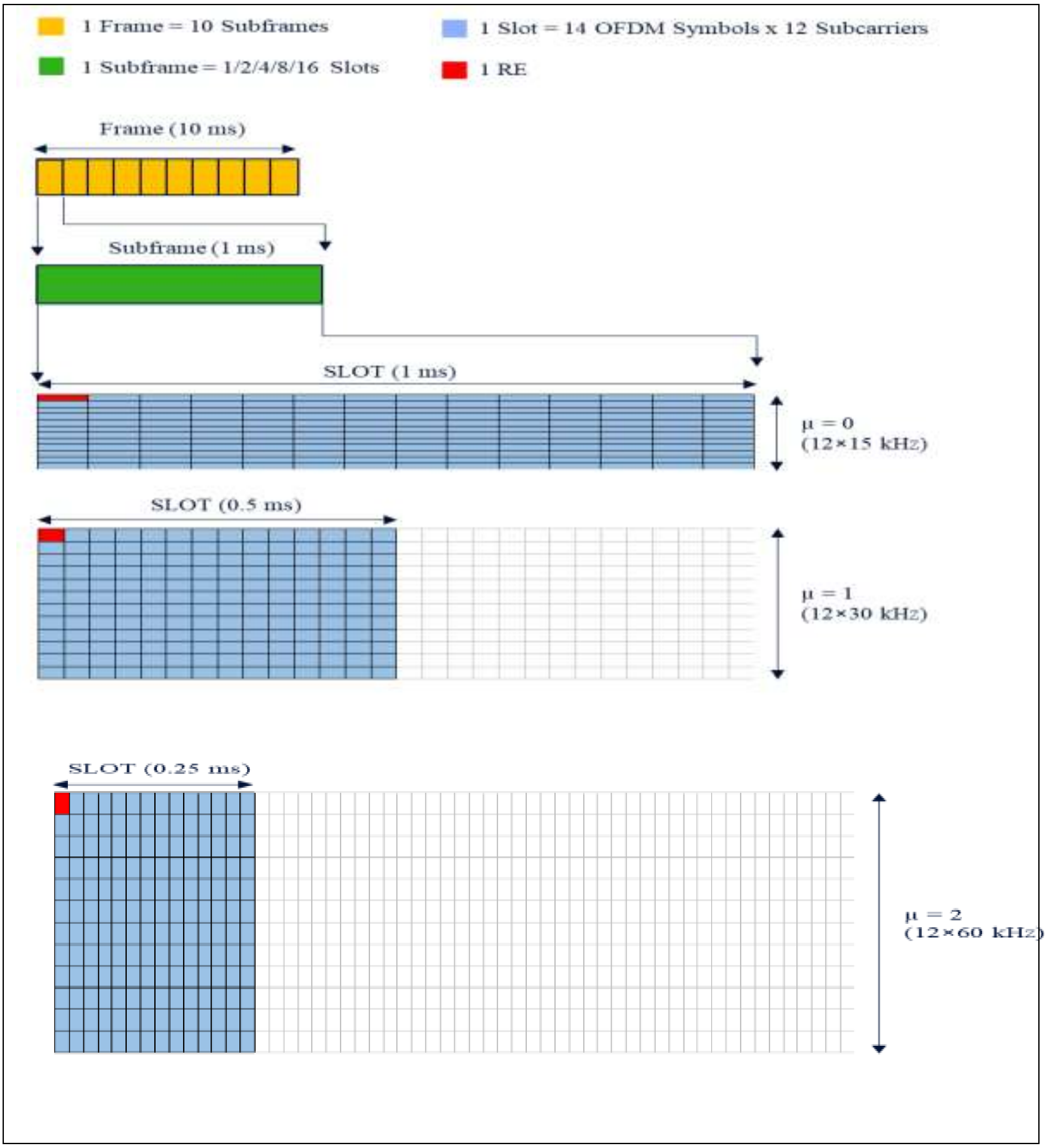


Figure III-9: NR Frame Structure

III-6-1 5G NR Physical Channels and Signals

Control and data content is allocated at the described 5G NR frame structure for downlink and uplink transmissions thanks to the use of physical channels and signals. Since the scope of this master thesis is mainly focused on the downlink side, only the physical channels and signals transmitted from the base station (NB) to the UE are described in this subsection. The set of downlink physical channels and signals is shown in Figure III-10 [37]

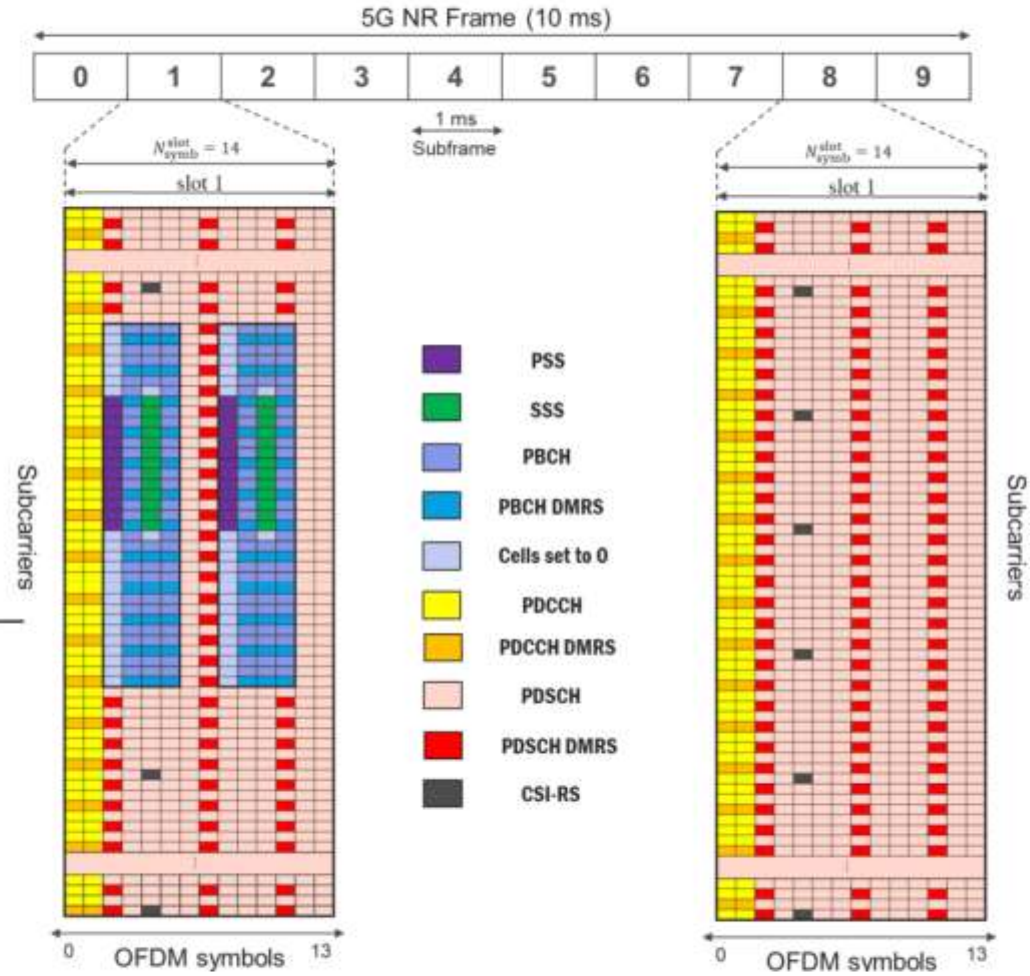


Figure III-10: Frame structure for SFI = 0 (10 subframe, 100% Downlink).

III-6-1-1 PHYSICAL SHARED CHANNELS

A. Physical Downlink Shared Channel: PDSCH is used for the transmission of DL user data, UE specific higher layer information, system information, and paging.

For transmission of a DL transport block (payload for physical layer), a transport block CRC is first appended to provide error detection, followed by a LDPC base graph selection.we can find LDPC base graphs that the NR supports, the first is optimized for small transport blocks and the second is basically used for larger once . Then the segmentation of the transport block into code blocks and code block CRC attachment are performed. Where Each code-block is encoded , secondly, The LDPC coded blocks are then separatly rate matched. Finally, code block concatenation is done in order to create a code word for the transmission on the PDSCH channel, it supports transmission of Up to 2 codewords simultaneously on the PDSCH.

The contents of each codeword are scrambled and modulated to generate a block of complex-valued modulation symbols. The symbols are mapped on up to 4 MIMO layers. A PDSCH can have two codewords to support up to 8-layer transmission. The layers are then mapped to the ports of antenna in a specified transparent manner ,which is basically a non-codebook based, then explaining how the beamforming or MIMO precoding operation is done depends on the transparent and the network implementation to the user end . For each of the ports ex : layers used for transmission of the PDSCH, the symbols are mapped to Resource blocks RBs .[38]

Physical layer processing for NR PDSCH is summarized in in the left part of Figure III-11

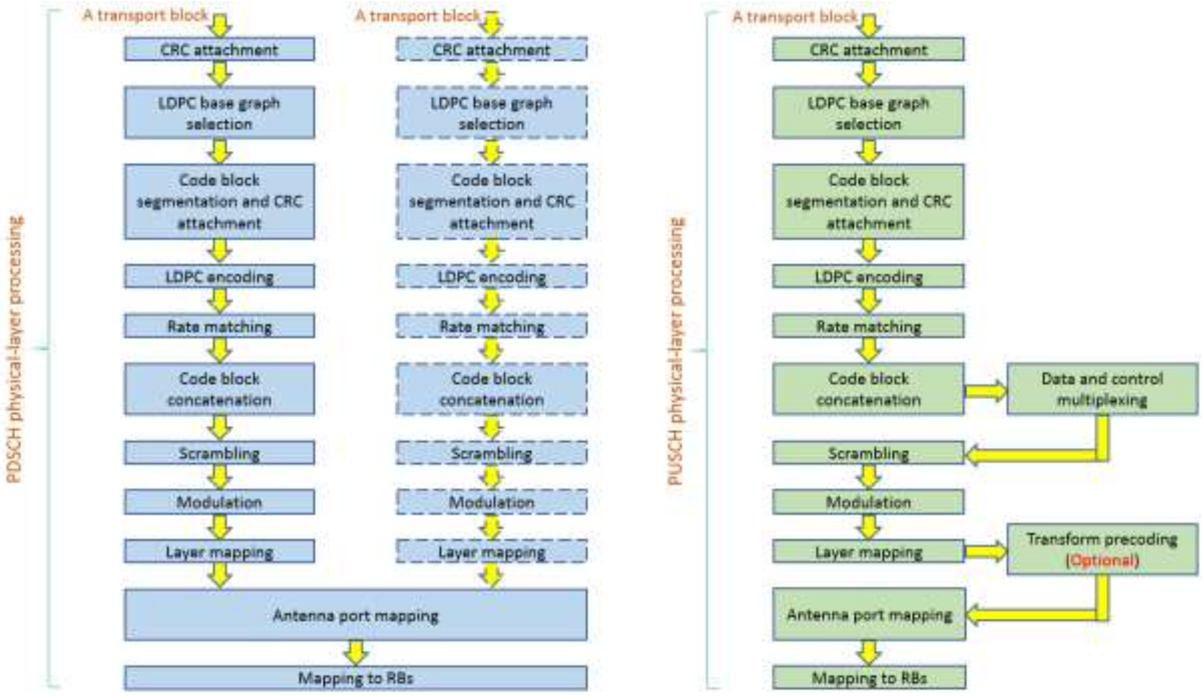


Figure III-11: Illustration of 5G NR PDSCH.

B. Physical uplink Shared Channel: PUSCH is used for the transmission of UL shared channel (UL-SCH) and layer 1/2 control information. The UL-SCH is the transport channel used for transmitting an UL transport block. The physical layer processing of an UL transport block is similar to the processing of a DL transport block, as summarized in right part of Figure III-11.[38]

The contents of the codeword are scrambled and modulated to generate a block of complex-valued modulation symbols. The symbols are then mapped onto one or several layers. PUSCH supports a single codeword that can be mapped up to 4 layers. In case of a single layer transmission only, DFT transform precoding can optionally be applied if enabled. For the layers to antenna ports mapping, both non-codebook-based transmission and codebook-based transmission are supported in the UL. For each of the antenna ports used for transmission of the physical channel, the symbols are mapped to RBs. In contrast to LTE, the mapping is done in frequency before time to enable early decoding at the receiver.

III-6-1-2 Physical Control Channels

A. Physical Downlink Control Channel: PDCCH is used to carry DCI such as downlink scheduling assignments and uplink scheduling grants. An illustration of NR PDCCH is given in the upper part of Figure III-12.

NR PDCCH channels are made in a special way to transmit in a configurable control resource set (CORESET). Where a CORESET is analogous to the control region in LTE , but is generalized in the terms that the group of RBs and the set of OFDM symbols in which it is located could be configurable with the corresponding PDCCH search spaces. this configuration flexibilities of the control regions may includes numerologies , frequency, time, and operating points ,which enable NR to address a various use cases.

A PDCCH is confined to one CORESET and transmitted with its own (DMRS) demodulation reference signal , this enables a specific UE beamforming of the control channel. A PDCCH is carried by 1, 2, 4, 8 or 16 control channel elements (CCE) to accommodate different DCI payload size or different coding rates. Each CCE consists of 6 REGs. The CCE-to-REG mapping for a CORESET can be interleaved (for frequency diversity) or non-interleaved (for localized beam-forming). A UE is configured to blindly monitor a number of PDCCH candidates of different DCI formats and different aggregation levels..[37]

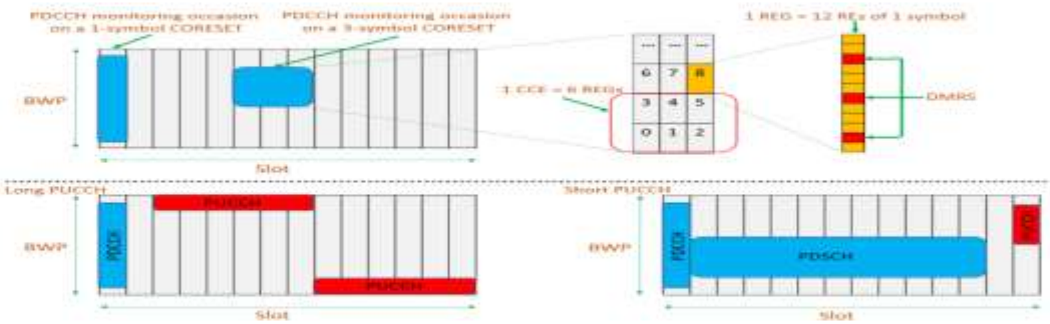


Figure III-12: Illustration of 5G NR PDCCH and PUCCH.

For the downlink, the modulation scheme is in this form [21]

Table III-6: Supported modulation schemes for downlink.

Modulation scheme	Modulation order Q_m
QPSK	2
16QAM	4
64QAM	6
256QAM	8

B. Physical Uplink Control Channel: PUCCH is used to carry uplink control information (UCI) such as scheduling request (SR). , channel state information (CSI), hybrid automatic repeat request (HARQ) feedback , and An illustration of NR PUCCH is shown in Figure III-12**,** Unlike LTE PUCCH that is located at the edges of the carrier, NR PUCCH is flexible in its time and frequency allocation. That allows supporting users who have a smaller bandwidth capabilities in an NR carrier and efficient usage of the resources with respect to capacity and coverage. NR PUCCH design is based on 5 formats of PUCCH. Short PUCCHs it uses formats (0 and 2) , use 1 or 2 OFDM symbols while for long PUCCHs it uses PUCCH formats (1, 3 ,4) and can use 4 to 14 OFDM symbols. PUCCH formats 0 and I carry UCI payloads of 1 or 2 bits while other formats are used for carrying UCI payloads of more than 2 bits. In PUCCH formats 1, 3 and 4, symbols with DMRS are time division multiplexed with UCI symbols to maintain low low peak-to-average-power-ratio (PAPR) while in format 2, DMRS is frequency-multiplexed with data-carrying subcarriers.

For the modulation in uplink.[21], the table illustrate ,in both precoding disabled and enabled.

Table III-7: Supported modulation schemes for uplink.

Transform precoding disabled		**Transform precoding enabled**	
Modulation scheme	**Modulation order** Q_m	**Modulation scheme**	**Modulation order** Q_m
		π/2-BPSK	1
QPSK	2	QPSK	2
16QAM	4	16QAM	4
64QAM	6	64QAM	6
256QAM	8	256QAM	8

III-6-1-3 PHYSICAL REFERENCE SIGNALS

Reference signal design follows the lean carrier principles -reference signals are on-demand when possible, and their time and frequency distributions are configurable so that requirements can be met with minimal overhead. To serve energy consumption reducing and intercell interference a t low load, the reference signal transmission can be extremely sparse.. the on-demand principle and the high flexibility of the design and also result in a degree of forward compatibly. While in LTE, multiple functions are tied up to the always-on CRS. In NR, these functions are supported by multiple users specifically configured reference signals[38]

A.Downlink and Uplink Demodulation Reference Signals (DMRS): DMRS is used by the receiver to produce channel estimates for demodulation of the associated physical channel. The design of DMRS is specific for cach physical channel - PBCH, PDCCH, PDSCH, PUSCH, and PUCCH. We can note that , a DMRS is specific for each UE , and only transmitted on demand, and in normal case it doesn't extend outside of the scheduled physical resource of the channel it supports. In the sequel, when CP-OFDM is used, we pay attention for PUSCH and the DMRS for PDSCH.

a wide range of scenarios is supported by The PDSCH/PUSCH DMRS, use cases, and user end capabilities, The number of DMRS symbols in a PDSCH and PUSCH duration can be configured; this enables support for very high UE mobilit ,but ,it also lower the DMRS overhead when the scenario allows so. In the same way , in the frequency domain and the density of DMRS is configurable to allow for an optimized overhead. The first DMRS instance comes early in the PDSCH or PUSCH transmission, this enables the receiver to start channel estimation to earlier, as consequence it reduces the processing latency. [38]

B. Downlink and Uplink Phase-Tracking Reference Signals(PTRS): tracking the phase of the local oscillator at the receiver and transmitter is done by using PTRS . This leads to eliminating the common phase error and phase noise, particularly important at high carrier frequencies , in the frequency domain PTRS can have low density but in the time domain it can have high density Due to the properties of phase noise, PTRS can be shown both in the in the uplink (PUSCH channel) and the downlink (PDSCH channel), where a PTRS is only associated with one DMRS port and is confined to the scheduled bandwidth and duration of PDSCH/PUSCH) If transmitted. The frequency and time densities of PTRS are adapted to scheduling bandwidth and signal-to-noise-ratio (SNR) [38]

C. Channel-State Information Reference Signals (CSI-RS): NR CSI-RS is used for down link CSI acquisition Similarly to the LTE counterpart,. Beyond this use case, CSI-RS in NR also could support reference signal received power (RSRP) measurements for beam management and mobility , frequency and time tracking for demodulation, and UL reciprocity-based precoding.

for CSI-RS configuration ,The NR supports high degree of flexibility. A resource can be configured with up to 32 ports, and the density is configurable. In the time domain, a CSI-RS resource may start at any OFDM symbol of a slot and it spans 1, 2, or 4 OFDM symbols depending on the number of ports configured. CSI-RS can be periodic, semi-persistent or aperiodic (DCI triggered). [37]

D. Sounding Reference Signals (SRS): For uplink channel sounding the SRS is used. The design supports uplink link schedulingand adaptation , but in reciprocity operation also downlink precoder selection, link scheduling and adaptation ,ex : for massive MU-MIMO.

As opposite to LTE, by making NR-SRS a user end specifically configured. It enables a high degree of flexibility in the system.

III-7 conclusion

After we have seen how frame structure could help to reach a low latency communication link E2E through several changes , this drives us to dig even deep more into how frame structure could be a very critical solution to latency in the next chapter , the evaluation mechanism using KPIs is very helpful to see and monitor the performance of each waveform and how it fits into the criteria of 5G NR ,although 5G is under tests and many candidates are proposed for being the principal waveform used for this great communication system , but what really makes the good choice is the one that fits most of 5G global three applications (eMBB,URLLC,mMTC),they researches event went further to suggest a non-orthogonal waveform due to its spectrum capabilities ,but until now it lacks reliability , 5G frame structure and channels focus more on the scalability because of the new concept of numerologies and the flexible time slot used in the frame .

Chapter IV: Simulation

IV-1 Introduction

5G frame structure ability to enable URLLC is due to the several modification in the physical layer ,one of the most significant modification is the frame structure flexibility as we have shown in chapter 03 and evaluated the waveform using the required KPIs for the work ,in the simulation we will use the MATLAB as its strong ability for studying signal processing as well as its library for 5G and also its accuracy in term of results ,in the first step we will simulate the burst time of most significant waveforms and see how it affects the latency ,in the second part we will focus on simulating the numerologies of 5G and how to manage high bandwidth decoding for small devices using the bandwidth-part technique

IV-2 simulation of burst duration for 5G candidates

As first part of this work, we selected CP-OFDM ,UFMC and FBMC to study the spectral efficiency depends on the burst time for the most highlighted 5G waveform candidates ,at we would have the next parameters :

Table IV -1: SIMULATION parameters for burst time duration

Bits per subcarrier	**6**
FFT size	**4096**
Number of cyclic prefix	**43**
Filter length	**43**

IV-2-1 Result and comments

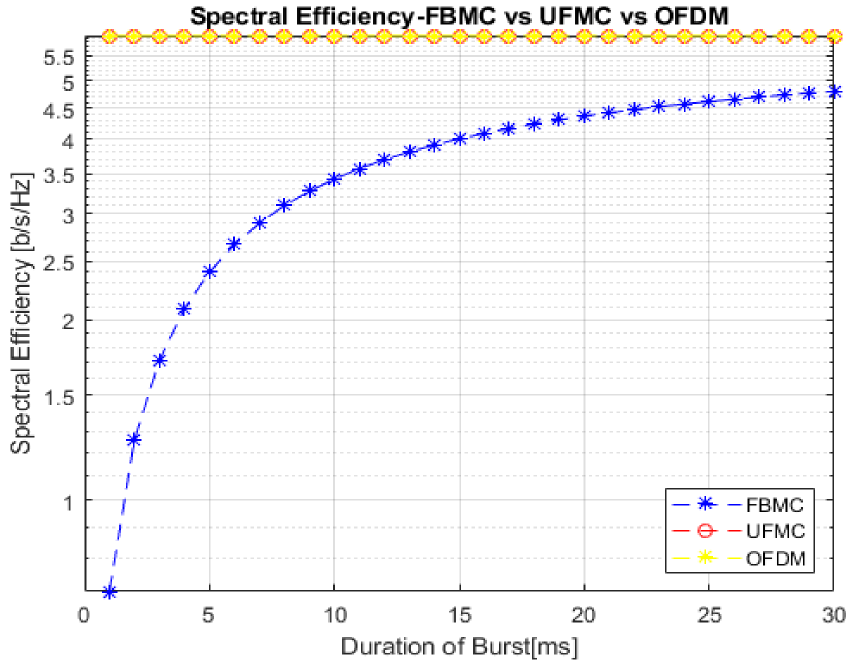


Figure IV-1: spectral efficiency and burst duration for 5G highlighted candidates

UFMC has an advantage compared to the OFDM by its higher spectral efficiency ,the ability of filtering the sub-bands has an impact on reducing the guard-bands between the sub-bands and thus it reduces the filter length , which makes it good for short bursts , while its also better than FBMC because the latter takes long time in the filtering as it filters every sub-carrier alone .

The graph above explains the spectral efficiency depends on the burst duration of the highlighted waveform candidates ,since we have both the CP-number and the Filter length the same to both OFDM and FBMC ,they overlap with each other over the burst, we also notice that FBMC increases its spectral efficiency by the increase of the burst duration , as it takes long burst for reaching a good spectral efficiency which makes it not suitable for low latency applications as it depends on the burst duration .

IV-3 simulation of frame structure of 5G

In the second phase of the simulation, our main focus is to see how 5G enables low latency with a reliability, this work could be separated to many parts as mentioned in chapter 01, RAN & CORE latency ,in our work ,we are going to simulate 5G frames and BWP for different cases to see how could the frame and bandwidth flexibility enable a big part of the latency .

Our simulation plan includes simulation of bandwidth and frame structure for a downlink with symbols distribution in different numerologies, we will also simulate EVM (error vector magnitude) of the NR test model for downlink, by the release 15,we have two frequency rages FR1 and FR2 ,

FR1 is between [50 MHz; 100 MHz] for frequencies lower than 6GHZ,

FR2 is between [100 MHz; 400 MHz] for mmwave frequencies

we are going to use NR-FR1-TM3.1 & NR-FR2-TM3.1 ,

Simulation parameters:

Table IV-2: SIMULATION parameters for frame structure

Frequency	FR1 < 6 Ghz		FR2 >> 6 Ghz (mmwave)	
MODULATION	QAM64	QAM64	QAM64	QAM64
Channel bandwidth	50 MHz	100 MHz	200 MHz	400 MHz
Subcarrier spacing (SCS)	15 KHz	30 KHz	60 KHz	120 KHz
Duplexing mode	TDD/FDD	TDD/FDD	TDD/FDD	TDD/FDD

CP-OFDM is used in 5G NR with variable subcarrier spacing so, it could reduce the symbol rate as well as benefits from the advantage of orthogonality which is reliability of transmission , In our result ,we have 8 frames in TDD since two are used for demodulator synchronization and 10 frames in FDD and we are going to simulate channels, BANDWIDTH PART.

IV-3-1 Channels

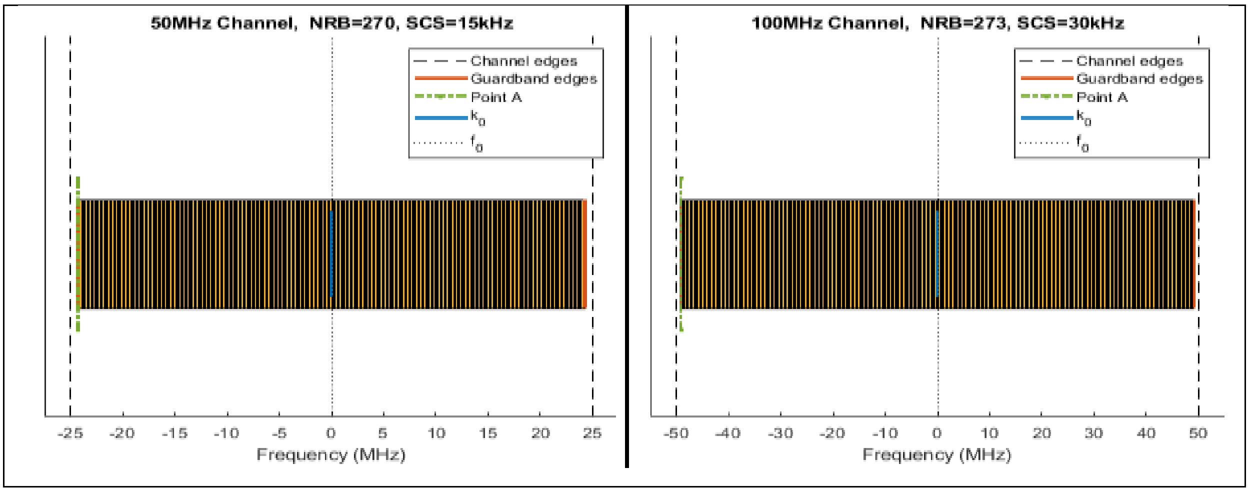


Figure IV-2: FR1 Downlink CHANNELS (TDD/FDD).

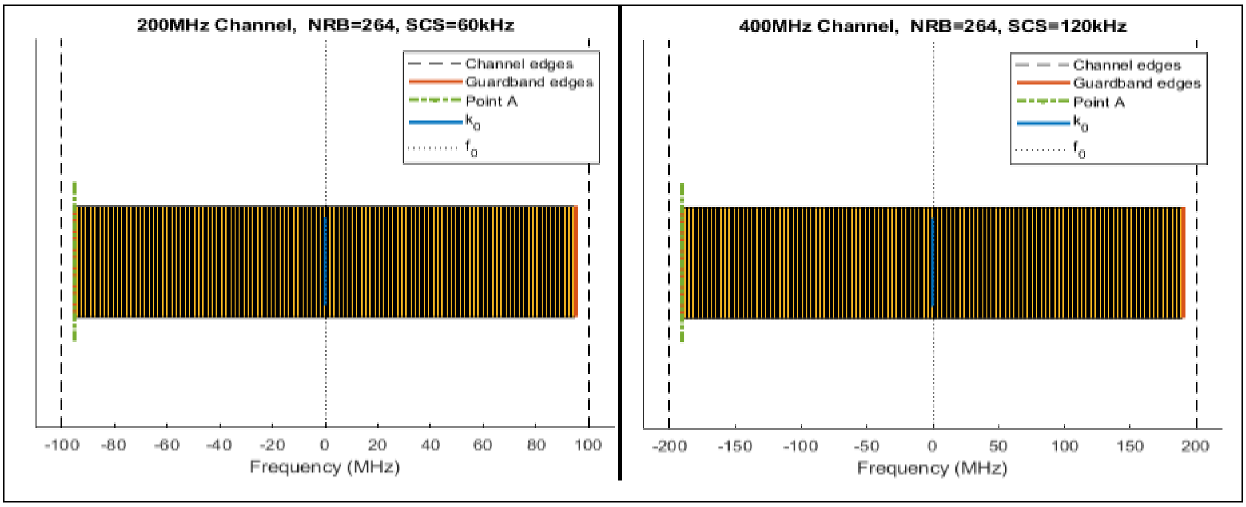


Figure IV-3: FR2 Downlink CHANNELS (TDD/FDD).

IV-3-1-1 Results and comments

In the result above , we have two kinds of channels ,mainly downlink channel ,FR1 belongs to the sub-6 Ghz , and the FR2 is for the mmwave channels bandwidth , as we can see , the maximum bandwidth allowed for FR1 is 100mhz ,where for FR2 we could reach up to 400mhz , this physical aspects made us dig more into the mapping of data and symbols in the channel .a guard band is always needed to make isolation ,while we also use channel edges ,mostly 20mhz , that introduces a new term called the nominal bandwidth or the usuable bandwidth for the data .

IV-3-2 TIME DIVISON DUPLEX (TDD)

IV-3-2-1 Bandwidth Parts (BWP)

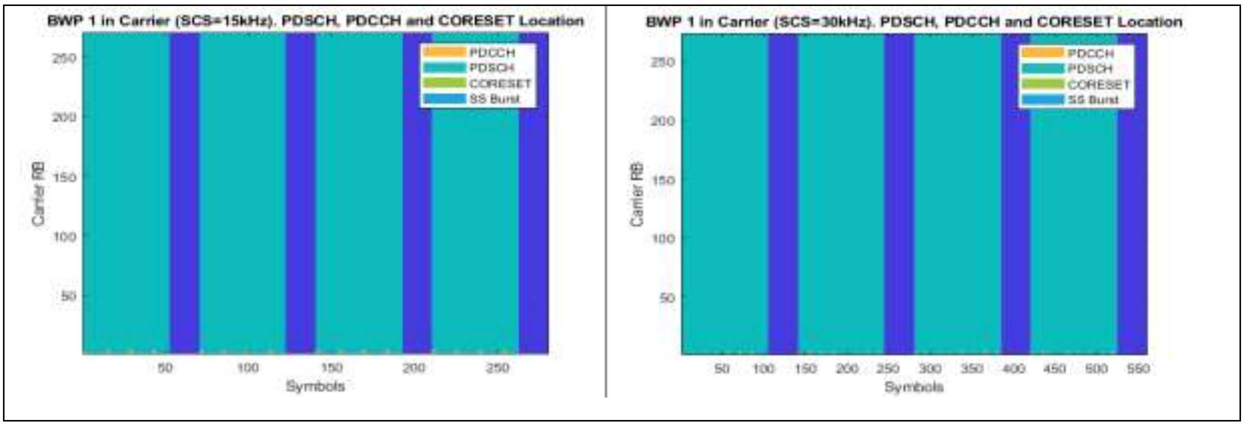


Figure IV-4: bandwidth part for FR1 subcarrier spacing (TDD).

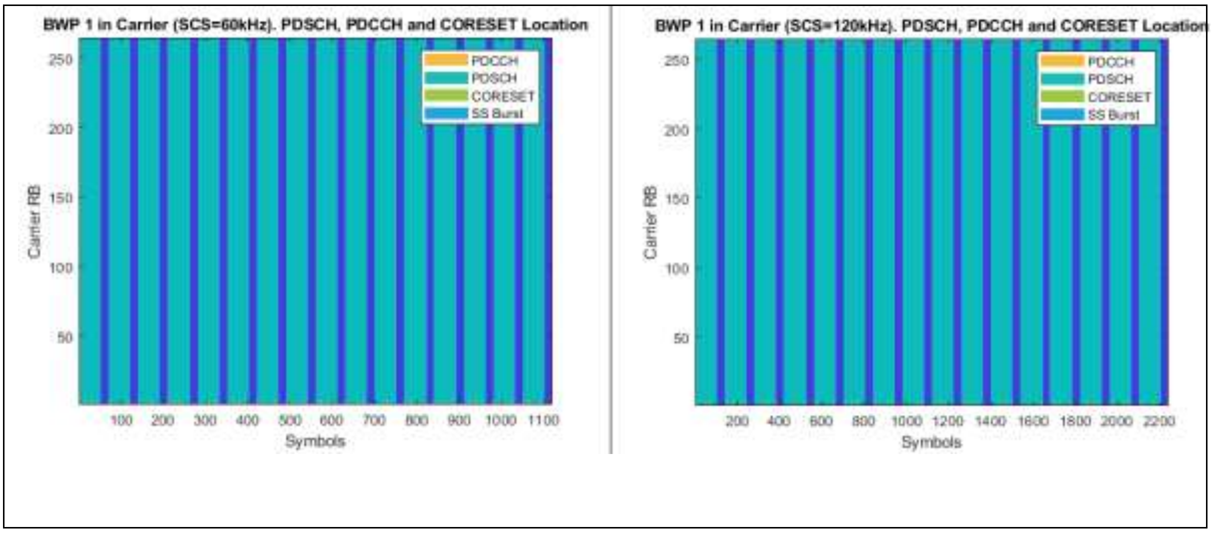


Figure IV-5: bandwidth part for FR2 subcarrier spacing (TDD).

IV-3-2-1-2 Results and comments

The specifications (TS 38.141-2) defines FR2 NR-TM for TDD only but this example also allows FDD waveforms to be created.NR has similar flexibility, with the ability to schedule as short as two symbol downlink (codewords)/one symbol uplink transmissions.

In 5G ,since we use much more higher bandwidth, UE doesn’t have to decode the whole bandwidth as in LTE , when UE is in idle mode ,he reads the information for other bandwidth parts (BWP) ,which exists in the initial bandwidth part.

For duplexing, FDD and TDD are two dominant duplex modes applied for paired spectrum and unpaired spectrum ,since TDD uses the whole burst at one time , we can dedicate a whole channel for one user ,this option made TDD very flexible with mmwave and large bandwdith compared to FDD , while TDD can be configured for both links subframe, but due to the requirement of synchronization , the configuration is kind relatively slow and the solution to address this problem ,we use a flexible duplex , this solution is suggested for 5G but not yet standardized .

For TDD, the subcarrier spacing, the monitoring capability, and the supported processing times. NR TDD with self-contained subframes can offer lower latency than the traditional TDD configurations supported in LTE, For the high frequency TDD carrier, the DL/UL time-slot (TS) ratio configuration only has to take into account the long-term DL/UL traffic statistics for guaranteeing the DL spectrum exploitation efficiency.

IV-3-2-2 Symbols/ Frames

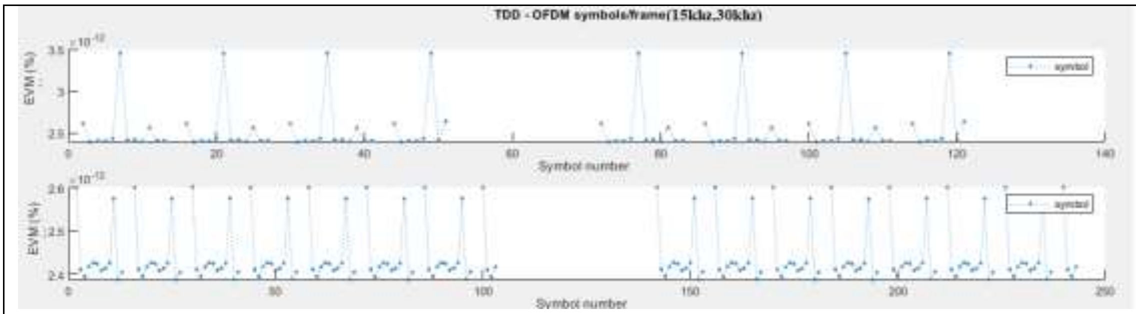


Figure IV-6: Frame simulation for FR1 (TDD)

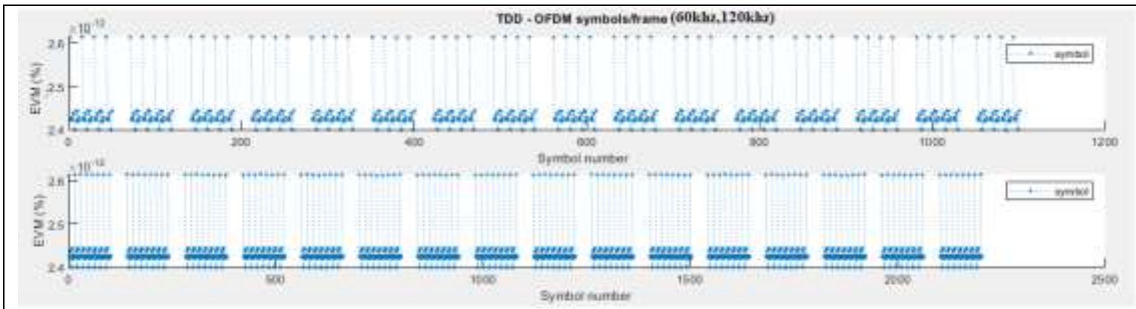


Figure IV-7: Frame simulation for FR2 (TDD)

IV-3-3 Frequency Division Duplex (FDD)

IV-3-3-1 Bandwidth Part (BWP)

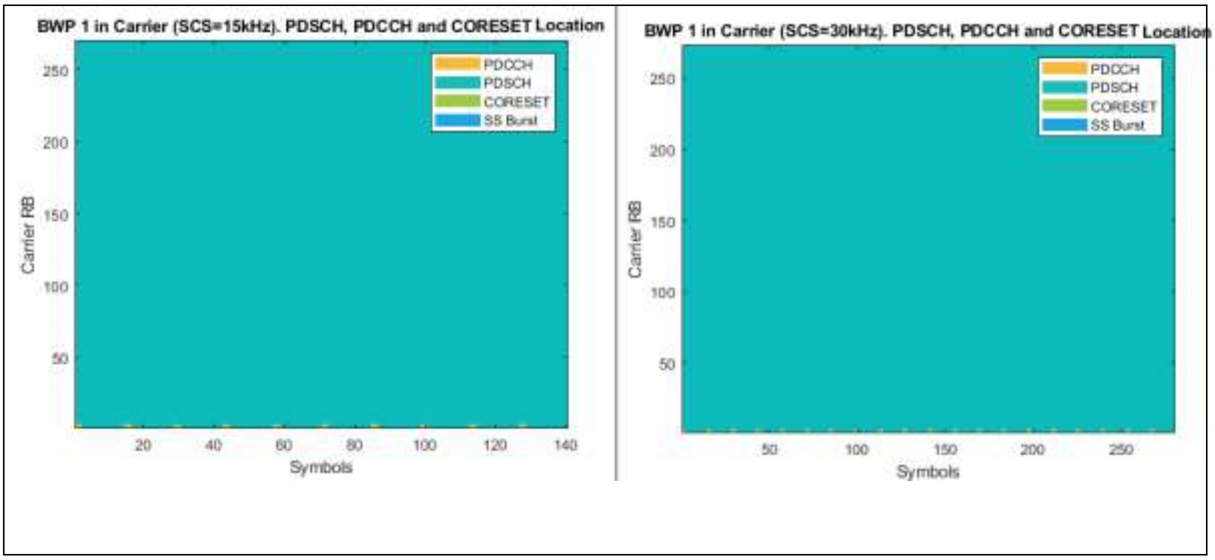


Figure IV-8: bandwidth part for FR1 subcarrier spacing (FDD)

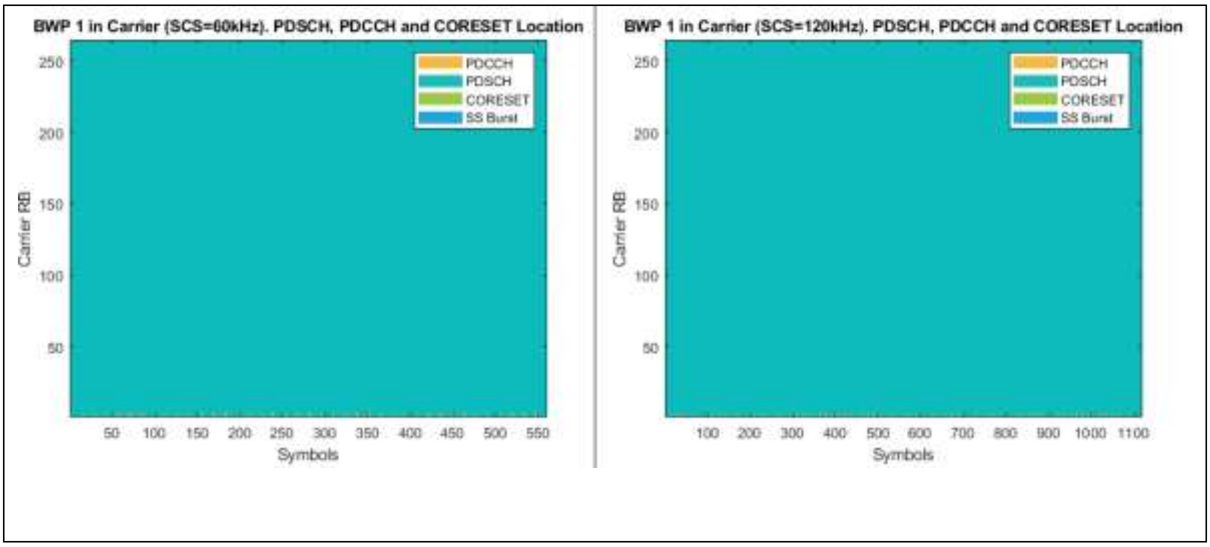


Figure IV-9: bandwidth part for FR2 subcarrier spacing (FDD)

IV-3-3-1-1 Results and comments

In the duplex mode , the Downlink channel could support both FDD and TDD , FDD could use two for FDD , Due to the more imbalance between the downlink and uplink load , the DL/UL symmetric bandwidths in FDD results an non-effective way of using the UL spectrum, Due to the flexible duplexing, there is a very flexible frame structure for NR and each symbol in one slot can be flexibly configured as downlink or uplink transmission in principle, With a larger possible subcarrier spacing, latency can, thusly, be significantly reduced. The performance of NR for URLLC will strongly depend on whether FDD or TDD is used and the configuration .

A UE can be configured with up to four bandwidth parts in the downlink, with a single downlink bandwidth part being active at one time within a particular SCS,this configuration is made for much more simplicity of implementation of this concept ,bwp is characterized by its SCS and cyclic prefix, Supporting reduced UE bandwidth capability is especially helpful for devices with limited RF capability or those not capable of full carrier bandwidth.as well as Supporting reduced UE power consumption for intermittent and bursty traffic profiles. and during the connected mode ,UE could change into different SCS's bwp .

Also ,in 5GNR , DC is modulated because in concept of bandwidth part ,each bwp needs to be centralized within its carrier frequency .

IV-3-3-2 Symbols/Frames

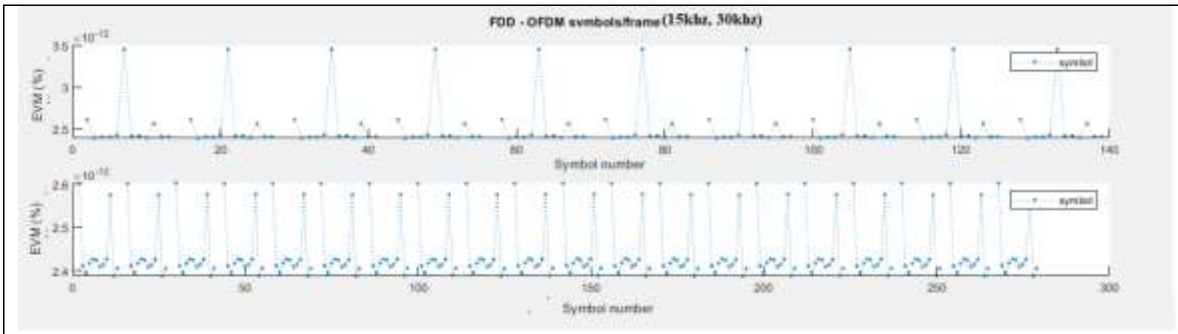


Figure IV-10: Frame simulation for FR1 (FDD)

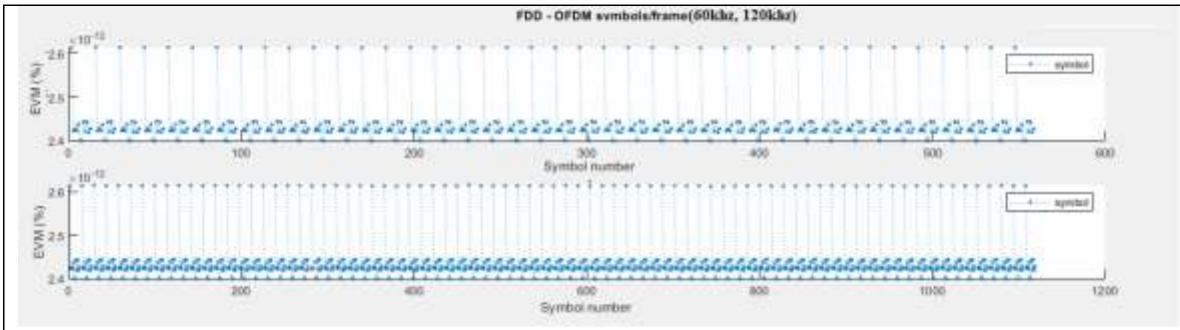


Figure IV-11: Frame simulation for FR2 (FDD)

IV-4 Result & Comments

5G NR supports flexible frame which leads to a flexible symbol duration According to the frame structure of 5G NR, 30 Khz SCS will reduce the symbol length by 50%, and 60 kHz SCS can reduce the length by 75%. This would help to reduce the latency of the steps related to air interface transmission for both user plane latency and control plane latency.we have different SCS for different use-cases .

We used Error Vector magnitude EVM measurement in one frame structure, without using any impairment. , (EVM) can be used as an effective and efficient method to evaluate the RF performance of those digitally modulated communication systems ,where The EVM is defined as the length of the vector between the received symbol and the transmitted symbol, For a correct demodulation it is necessary to know which symbol has been sent. Both LTE and 5GNR assume in their EVM measurement algorithms that the transmitted symbol is the one which is next to the received one.

Flexible frame duration is due to flexible subframe structure ,where each subframe has number of slots , slots in 5GNR are not same as in LTE (in LTE slots are fixed in time to 1ms) ,in this case ,frame could take wide different number of symbols in a fixed frame duration because of the wider subcarrier spacing ,which leads to lower symbol duration

Frame duration in 5GNR is 10ms, where each subframe duration is 1ms, and where slots number in each subframe is flexible which leads to flexible number of symbols, which means flexible symbol duration

as mentioned in Figure IV-10 & Figure IV-11 as well as Figure IV-6 & Figure IV-7

- 15KHZ SCS : we have one slot per subframe ,each slot has 14 OFDM symbols , duration of 10 slots is 10 ms,(slot duration is 1ms).

- 30KHz SCS : we have two slots per subframe ,each slot has 14 OFDM symbols ,duration of 20 slots is 10ms ,(slot duration is 0.5ms).

- 60KHz SCS : ,we have two slots per subframe ,each slot has 14 OFDM symbols ,duration of 40 slots is 10ms ,(slot duration is 0.25 ms).

-120KHz SCS : we have two slots per sub frame ,each slot has 14 OFDM symbols ,duration of 80 slots is 10ms ,(slot duration is 0.125ms).

after this result, we can see that the flexible subframe format leads to a very low symbol duration which enables low latency, as time critical applications ,we can use wider SCS for users to provide them with a very low latency symbol duration.

In addition , we notice that in FR2 ,we use mmwave and 60khz is kind of high frequency for the users to deal with,

as solution to this ,5G NR proposes Bandwidth part ,which enables user to decode only a part for the bandwidth as mentioned in Figure IV-4 & 5 ,as well as Figure IV-8 & 9, the difference is in TDD the simulator simulates only 8 frames, and two others are for synchronization.

Mini-slot (2,4, or 7 symbols) can be allocated for shorter transmissions, where Slots can also be aggregated for longer transmissions

Note: we can notice there are only 12 symbols ,the other two are suppressed for the cyclic prefix(usually first and last one)

IV-4-4 Other Interesting Facts about 5GNR

NR channel coding is different from other generations, we use both LDPC power for high data rates and use polar code for the short transmissions , as LDPC has an advantage of decoding latency and the polar code shows the performance advantage when it is used for control channel.

IV-5 Conclusion

As we have seen so far , the frame structure for downlink has many parameters ,starting from the channel shape to the symbols distribution in the frame , while we use fixed slot time and different number of symbols depending on the bandwidth (subcarrier spacing) we use ,by the evaluation of the top three waveforms candidates used for 5G , Each of the multiple access techniques has its own advantage but they lack flexibility, for ex : (FBMC) is good for big packet but doesn't handle latency in small packets, (NOMA) is good for latency but it’s not reliable,

Reliability is also related to latency, it’s defined as if the one time transmission latency can be reduced, and there can be opportunity to have more repetitions within the latency budget. The reliability can be therefore improved.

For release 15 and for NR new physical layer ,CP-OFDM has been chosen to be the waveform for both the links ,while a small modification has been done in the OFDM by introducing the DFT-S-OFDM in the uplink for its power of coverage as well as low energy consumption for the end user devices ,the choice of the CP-OFDM is the most optimum solutions after several tests between the waveform candidates , the use of cyclic prefix lets the use on one tap equalization in the frequency domain after the modulation , also , it helps in the robustness against multipath problems due to the duration ,as wel as its much more high spectral efficient and low complex in term of cost

The deployment of 5G spectrum varies from country to other one ,but the standards which are introduced by the 3GPP we can obtain a 100mhz bandwidth easily in the sub-6GHZ frequency ,which is an efficient bandwidth compared to the 4G which has a 20mhz maximum bandwidth , also ,the use of multiple numerologies enables different scenarios and applications such as URLLC ,eMBB,mMTC to be deployed . narrow subcarriers allows the user for less battery conception and good in idle mode (15khz for ex) and in active mode we assign the user with wider SCS and as a result short symbol duration.

General Conclusion

5G requirements are still a promise in 2020, reaching all the requirements necessity is still a long road, for 1ms total latency and 10Gbps, as well as the deployment of the 5G around the globe.

Fulfilling the world needs in the telecom in general or in the cellular network in particular is based on valuing the work and the collaboration of each individual person, a small modification or optimization whether in latency optimization or in throughput or any entity could really makes the different in reaching the theoretical requirement of 5G system, we believe that the key to innovation journey starts with understanding the technology, in our work we focused on giving the most details of how latency works and the significant aspects of it, from the RAN to the transport as explained in the chapter three, the latency is a divided to many section each part has the necessity to be optimized as well as making it much more reliable, many technologies are now possible due to the other works in the different industries, such as LDPC coding which was an old system but glad to the hardware technology ,now its possible to use its features in the 5G ,same goes for the beamforming which was already discovered in the analog and have been used in many old technologies, also, SDN and NFV are now possible due to the vast capacity and processing speed, which made the software tasks as fast as hardware, with this concepts, we no longer have to make Application specific integrated-circuits (ASIC) and throw / recycle them after few years, the software made it easier for us to build any type of network architecture and monitor/control our system with much more precision with a lower cost.

The frame structure has a significant impact on the latency in 5G was brought by the flows made in the 4G as well as the new capability of utilizing the resources we have, such as mmwave massive bandwidth, the idea of variable numerologies made its fingerprint on the latency aspect, since there is a reciprocal relation between high frequencies and thus made the symbol duration smaller. However, due to higher frequencies we face a lot of other problems especially the possibility for an IoT or low power device to decode the whole bandwidth, here where it comes the idea of the Bandwidth part, the UE receives only a part of the bandwidth then decode it with ease, another aspect is the mini-slot, where we send 2,4 or 7 symbols for a user rather than 14 OFDM symbol, this would help the applications which depends on latency more than the throughput.

Although this subject is still under research and many suggestions could be done to optimize the total latency in 5G system, many types of research are targeting scheduling techniques for the frame structure in RAN and the cloud scheduling algorithms in CORE NETWORK, caching helps to minimize the route from UE to cloud which makes this idea also easier to be applicable and handy.

We suggest that the next contribution would be done on the core network latency or caching ideas for much more optimization and understanding of the latency aspect.

References

[1] Alex Barakabitze, Md. A. A. (2015). Evolution of LTE and Related Technologies towards IMT-Advanced. Evolution of LTE and Related Technologies towards IMT-Advanced, 1-4.

[2] Holma, H., & Toskala, A. (2012). LTE Advanced. Hoboken, NJ, États-Unis : Wiley.

[3] Vaezi, M., Ding, Z., & Poor, H. V. (Éd.). (2019). Multiple Access Techniques for 5G Wireless Networks and Beyond.

[4] Hoymann, C., Astely, D., Stattin, M., Wikstrom, G., Cheng, J.-F., Hoglund, A.,… Gunnarsson, F. (2016). LTE release 14 outlook. IEEE Communications Magazine, 54(6), 44-49. https://doi.org/10.1109/mcom.2016.7497765

[5] Ma, Z., Zhang, Z., Ding, Z., Fan, P., & Li, H. (2015). Key techniques for 5G wireless communications: network architecture, physical layer, and MAC layer perspectives. Science China Information Sciences, 58(4), 1-20. https://doi.org/10.1007/s11432-015-5293-y

[6] Dahlman, E., Parkvall, S., & Skold, J. (2018). 5G NR : The Next Generation Wireless Access Technology (1re éd.). chennai, india : Academic Press.

[07] Holma, H., Toskala, A., & Nakamura, T. (2019). 5G Technology. Hoboken, NJ, États-Unis : Wiley.

[08] 3GPP releases 16 & 17 overview – the 5G NR evolution. (s. d.). Consulté le 15 juillet 2020, à l'adresse https://www.ericsson.com/en/reports-and-papers/ericsson-technology-review/articles/5g-nr-evolution#:%7E:text=16%20%26%2017%20overview-,5G%20evolution%3A%203GPP%20releases%2016%20%26%2017%20overview,new%20verticals%20and%20deployment%20scenarios.

[09] Yilmaz, O. (s. d.). Key innovation areas of 3GPP Rel-17. Consulté le 15 juillet 2020, à l'adresse https://www.ericsson.com/en/blog/2019/12/3gpp-rel-17

[10] Dahlman, E., Parkvall, S., & Skold, J. (2018). 5G NR : The Next Generation Wireless Access Technology (1re éd.). chennai, india : Academic Press.

[11] 3GPP Release 15 Overview. (s. d.). Consulté le 15 juillet 2020, à l'adresse https://spectrum.ieee.org/telecom/wireless/3gpp-release-15-overview

[12] Nguyen, V.-G., Do, T.-X., & Kim, Y. (2015). SDN and Virtualization-Based LTE Mobile Network Architectures : A Comprehensive Survey. Wireless Personal Communications, 86(3), 1401-1438. https://doi.org/10.1007/s11277-015-2997-7

[13] Meerasri, P., Uthansakul, P., & Uthansakul, M. (2014). Self-Interference Cancellation-Based Mutual-Coupling Model for Full-Duplex Single-Channel MIMO Systems. International Journal of Antennas and Propagation, 2014, 1-10. https://doi.org/10.1155/2014/405487

[14] Liyanage, M., Ahmad, I., Abro, A. B., Gurtov, A., & Ylianttila, M. (2018). A Comprehensive Guide to 5G Security (1re éd.). finland, finland : Wiley.

[15] Wei L L, Hu R Q, Qian Y, et al. Enable device-to-device communications underlaying cellular networks: challenges and research aspects. IEEE Commun Mag, 2014, 52: 90–96

[16] Kim, K.-J., Hong Kang, J., Hwang, J.-H., & Ahn, K.-H. (2018). Hybrid Beamforming Architecture and Wide Bandwidth True-Time Delay for Future High Speed Communications 5G and Beyond 5G Beamforming System. 2018 IEEE 3rd International Conference on Integrated Circuits and Microsystems (ICICM), 1-3

[17] Beamforming management and beam training. (2019). Ahmed, Samir, 1-16. Consulté à l'adresse https://trepo.tuni.fi/handle/10024/118195

[18] Ekström, H. (2019, juillet 7). Non-standalone and Standalone: two paths to 5G. Consulté à l'adresse https://www.ericsson.com/en/blog/2019/7/standalone-and-non-standalone-5g-nr-two-5g-tracks , at 15:05 , 25/04/2020

[19] Bennis, M., Debbah, M., & Poor, H. V. (2018). Ultrareliable and Low-Latency Wireless Communication : Tail, Risk, and Scale. Proceedings of the IEEE, 106(10), 1834-1853. https://doi.org/10.1109/jproc.2018.2867029

[20] Lei, Wan & Soong, Anthony & Jianghua, Liu & Yong, Wu & Classon, Brian & Xiao, Weimin & Mazzarese, David & Yang, Zhao & Saboorian, Tony. (2020). 5G System Design: An End to End Perspective. 10.1007/978-3-030-22236-9.

[21] 3GPP TR38.913, "Study on Scenarios and Requirements for Next Generation Access Technologies", June 2017

[22] Parvez, I., Rahmati, A., Guvenc, I., Sarwat, A. I., & Dai, H. (2018). A Survey on Low Latency Towards 5G: RAN, Core Network and Caching Solutions. IEEE Communications Surveys & Tutorials, 20(4), 3098-3130.

[23] She, C., Yang, C., & Quek, T. Q. S. (2017). Radio Resource Management for Ultra-Reliable and Low-Latency Communications. IEEE Communications Magazine, 55(6), 72-78. https://doi.org/10.1109/mcom.2017.1601092

[24] Hyoungju Ji, Sunho Park. (2018). Ultra Reliable and Low Latency Communications in 5G Downlink: Physical Layer Aspects. Samsung Electronics, Korea. Dept. of Electrical and Computer Engineering, Seoul National University, Korea Samsung Electronics, Korea

[25] Radiocommunication sector of ITU. (2017). Minimum requirements related to technical performance for IMT-2020 radio interface(s) (Report ITU-R M.2410-0). Adresse : http://www.itu.int/ITU-R/go/patents/en

[26] I.-G. Jang and G.-D. Jo, "Study on the latency efficient IFFT design method for low latency communication systems," in Proc. Int. Symp.Intell. Signal Process. Commun. Syst. (ISPACS), Oct. 2016

[27] Dutta, Sourjya & Mezzavilla, Marco & Ford, Russell & Zhang, Menglei & Rangan, Sundeep & Zorzi, Michele. (2016). MAC layer frame design for millimeter wave cellular system. 117-121. 10.1109/EuCNC.2016.7561016.

[28] 3GPP Technical Specifications 38.211, "Technical Specification Group Radio Access Network, NR (Release 15)," v15.0.0,2017.

[29] R. Gerzaguet, N. Bartzoudis, L.G. Baltar, V. Berg, J.B. Doré, D. Kténas, O. Font-Bach, X. Mestre, M. Payaró, M. Färber, K. Roth, The 5G candidate waveform race: a comparison of complexity and performance, EURASIP Journal on Wireless Communications and Networking 2017 (1) (2017, Jan.).

[30] Zaidi, A., Athley, F., Medbo, J., Durisi, G., Gustavsson, U., & Chen, X. (2018). 5G Physical Layer. Maarssen, Pays-Bas : Elsevier Gezondheidszorg

[31] Vaezi, M., Ding, Z., & Poor, H. V. (Éd.). (2019). Multiple Access Techniques for 5G Wireless Networks and Beyond. *Multiple Access Techniques for 5G Wireless Networks and Beyond*, https://doi.org/10.1007/978-3-319-92090-0

[32] Dr. Essam Sourour, P. S. B. A. U. (2019). Non Orthogona lMultiple Access as a Candidate for 5G (1). Consult the address https://engw.psau.edu.sa/sites/default/files/field/attached/Seminar%20Presentation-Prof%20Dr%20Essam%20Sourour%20on%20NOMA.pdf at 15:14 , 05/05/2020

[33] NOMA vs OMA | Difference between NOMA and OMA | 5G. (s. d.). Consulted in 1 mai 2020, address https://www.rfwireless-world.com/Terminology/NOMA-vs-OMA.html 16:10 , 05/05/2020

[34] Dahlman E., Parkvall S., Sköld J (2018) 5G NR: The Next Generation Wireless Access Technology. Academic Press, 466 p.

[35] Juho Lahdenperä. (2019). FEASIBILITY OF THE UE POWER SAVING SIGNAL FOR THE 5G NEW RADIO. Consulté à l'adresse https://www.semanticscholar.org/paper/On-feasibility-of-the-UE-power-saving-signal-for-5G-Lahdenper%C3%A4/a3b3b87a97ef250ce9c45cbd1f87940e178852d4

[36] Zaidi A., Baldemair R., Andersson M., Faxér S., Molés-Cases V. & Wang Z. (2017) 5G New Radio: Designing for the future, Ericsson technology Review. Ericsson AB, Sweden, Stockholm.

[37] José Luis Cárcel Cervera. (2018b). DESIGN AND PERFORMANCE EVALUATION OF A 5G NEW RADIO BROADCAST MODE FOR SINGLE FREQUENCY NETWORKS. Consulté à l'adresse https://pdfs.semanticscholar.org/70e8/c3857f7996456da5665d5cd83e769448f8e2.pdf

[38] Lin, X., Grovlen, A., Werner, K., Li, J., Baldemair, R., Cheng, J.-F. T., … Falahati, S. (2019b). 5G New Radio: Unveiling the Essentials of the Next Generation Wireless Access Technology. IEEE Communications Standards Magazine, 3(3). https://doi.org/10.1109/mcomstd.001.1800036