Mobile and wireless communications Enablers for the Twenty-twenty Information Society-II

Deliverable D4.2
Final air interface harmonization and user plane design

Version: v1.0

2017-04-30

http://www.5g-ppp.eu/

Deliverable D4.2
Final air interface harmonization and user plane design

Grant Agreement Number:	671680
Project Name:	Mobile and wireless communications Enablers for the Twenty-twenty Information Society-II
Project Acronym:	METIS-II
Document Number:	METIS-II/D4.2
Document Title:	Final air interface harmonization and user plane design
Version:	v1.0
Delivery Date:	2017-04-30
Editor(s):	Jakob Belschner (Deutsche Telekom AG); Emmanouil Pateromichelakis (Huawei Technologies GRC); Tomasz Mach (Samsung); Daniel Calabuig (Universitat Politècnica de València).
Authors:	Jakob Belschner, Gerd Zimmermann (Deutsche Telekom AG); Jamal Bazzi (DOCOMO Euro-Labs); Caner Kilinc, Ali Zaidi (Ericsson); Emmanouil Pateromichelakis, Malte Schellmann (Huawei Technologies GRC); Miltiadis Filippou (Intel Deutschland GmbH); Jens Gebert, Venkatkumar Venkatasubramanian (Nokia Bell Labs); Tomasz Mach, Yinan Qi, Milos Tesanovic, Wei Guo (Samsung); Nandish P. Kuruvatti (Universität Kaiserslautern); Daniel Calabuig, Sandra Roger, David Garcia-Roger (Universitat Politècnica de València).
Keywords:	air interface design, multi-connectivity, multi-service support, cellular protocol stack, 5G waveforms, user plane design

Status:	Final
Dissemination level:	Public

Abstract

This deliverable provides a harmonized 5th Generation (5G) Air Interface (AI) design that enables support of a large versatile type of services with different and often diverging requirements. The physical (PHY) layer issues relevant to harmonization are addressed and evaluated against harmonization Key Performance Indicators (KPIs) developed in METIS-II D4.1. Furthermore, the service-tailored 5G User Plane (UP) architecture design is investigated considering extreme Mobile Broadband (xMBB), massive Machine Type Communications (mMTC), and ultra-reliable Machine Type Communication (uMTC) aka Ultra-Reliable Low Latency Communications (URLLC). Finally, UP design impacts on the overall Radio Access Network (RAN) architecture are studied with focus on Control Plane (CP)/ UP functional split and network slicing.

Revision History

Revision	Date	Description
0.1	2017-03-08	Included Chapter 2, Annexes A and B, and references and abbreviations therein.
0.2	2017-03-14	Included remaining chapters.
0.3	2017-03-17	Merged comments by Tomasz and Jens
0.4	2017-03-21	Addressed comments by Tomasz and Jens
0.5	2017-03-31	Addressed comments by Ji and David
1.0	2017-04-28	Final version including review comments

Executive summary

The 5G mobile communication paradigm aims to support a large versatile type of services with different and often diverging requirements, which has posed significant challenges on the design of 5G systems. The main objective of this deliverable is to address some of these challenges and provide a harmonized 5G air interface and user plane design that enables high performance and a flexible system architecture.

Firstly, the physical layer issues relevant to harmonization are addressed and evaluated against harmonization KPIs developed as part of METIS-II Work Package 4 work in D4.1. As one of the key physical layer components that determine the system KPIs such as throughput, reliability, and complexity, waveform harmonization is studied and two options are proposed: single waveform and multiple waveform, where in the latter option different waveforms, e.g., OFDM and FBMC, are harmonized. An all-encompassing solution in a single harmonized implementation is proposed and complexity and latency are analysed. The operating spectrum bands of 5G systems are expected to expand across a much wider range from 600 MHz to up to 100 GHz. In this regard, multiple bands harmonization is proposed and different types of services are mapped to different spectrum bands according to the service requirements and properties of the spectrum bands. For the case of multiple services co-existing on the same band, different numerologies and frame structures are proposed and the latest development in standardization organizations, e.g., 3rd Generation Partnership Project (3GPP), is summarized. Moreover, the impacts of harmonization to upper layers are discussed.

Secondly, the service-tailored 5G UP architecture design is investigated considering all 5G use cases like extreme Mobile Broadband, massive Machine Type Communications, and ultra-reliable Machine Type Communication. As a reference, Long-Term Evolution (LTE) air interface UP protocols and functions are reviewed and new challenges for 5G are identified. With harmonized multiple air interface variants (AIVs), key control functionalities such as AIV to resource mapping are needed by UP of New Radio (NR). Enhancements to existing protocol stacks and NR processes, e.g., Hybrid Automatic Repeat Request (HARQ) process, are proposed. The interaction and interface design between RAN and Core Network (CN) are investigated with focus on Quality of Service (QoS) models and comparison of different protocol options. The current 3GPP standardization status summary on 5G UP design is also provided.

Finally, UP design impacts on the overall RAN architecture are studied with focus on CP/UP functional split and network slicing. Two scenarios are considered including 5G AIVs only and integration of 5G with legacy networks, e.g., LTE, and a service-based functional split to effectively aggregate multiple services while maintaining service-specific configurability. Network slicing, proposed to simultaneously address the potentially diverging requirements in terms of the latency, data rate and reliability and its impact on the CP/UP functional split is also studied.

Contents

1 Introduction 11
1.1 Objective of the document 11
1.2 Structure of the document 12
2 PHY layer aspects that enable multi-service support 13
2.1 Introduction 13
2.2 Multi-waveform harmonization 14
2.2.1 3GPP waveforms study work 15
2.2.2 Background to multi-waveform usage 15
2.2.3 Multi-waveform implementation proposal 16
2.2.4 Investigating the latency performance of in-band multi-waveform harmonization 24
2.3 Considerations about the co-existence of multiple bands 26
2.4 Considerations about the co-existence of different numerologies and frame structures in the same band 28
2.4.1 Different numerologies in the same band 28
2.4.2 Different frame structures in the same band 35
2.5 Harmonized coherent and non-coherent reception 37
2.5.1 Introduction and motivation 37
2.5.2 Constellation structure 38
2.5.3 Equivalent constellations 38
2.5.4 Harmonized decoders and performance evaluation 39
2.6 Considerations for higher layers 40
2.7 Conclusions on PHY layer aspects enabling multi-service support 41
3 Service-oriented functional UP design 43
3.1 Air interface UP Protocols and Functions: Overview of main challenges 44
3.2 Air Interface protocol configuration and functional mapping for multi-service support . 45
3.2.1 Requirement for novel NR Functions 46
3.2.2 Enhancements to existing protocol stack 47
3.2.3 Enhancements to NR processes 50
3.3 RAN/CN Design Considerations 64
3.3.1 5G QoS model considerations 64

3.3.2 Thin pipes vs. fat pipes on the RAN-Core interface 67
3.3.3 Protocol options for the RAN-CN interface 68
3.4 5G Standardization Status on UP Design 72
3.5 Conclusions on Service-oriented functional user plane design 74
4 User Plane design considerations on overall RAN architecture 75
4.1 Split of RAN user plane functions in central and distributed units 75
4.1.1 Functional split within a 5G air interface variant 76
4.1.2 Functional split considerations including eLTE 79
4.1.3 Service-based Functional Split / Placement Considerations 80
4.2 Separation of RAN control plane functions from user plane functions 83
4.2.1 Motivation 83
4.2.2 CP/UP split based network architecture 84
4.2.3 CP/UP-split variations in RAN deployments 88
4.3 Impact of Network Slicing on the User Plane 89
4.3.1 Network slicing in a 5G RAN 89
4.3.2 Network Slicing impact on CU-DU functional split 90
4.4 Conclusions on user plane design considerations on overall RAN architecture 91
5 Conclusions 93
6 References 96
A Further implementation aspects 102
A.1 Fast Fourier transform implementation 102
A.2 DFT spreading implementation 102
A.3 Filter bank implementation through a polyphaser network 103
B Single-waveform implementation complexity 105
B.1 Complexity of CP-OFDM 105
B.2 Complexity of W-OFDM 105
B.3 Complexity of ZT-DFT-s-OFDM 105
B.4 Complexity of P-OFDM 106
B.5 Complexity of FBMC-QAM 106
B.6 Complexity of FBMC-OQAM 106
C Assumptions for Calculations of Functional Split Data Rates 108

List of Abbreviations and Acronyms

3GPP	Third Generation Partnership Project
5G-PPP	5G Private Public Partnership
ACK	Acknowledgement
ACLR	Adjacent Channel Leakage Ratio
AFB	Analysis Filter Bank
AI	Air Interface
AIV	Air Interface Variant
AM	Acknowledged Mode
AP	Access Point
APN	Access Point Name
AR	Augmented Reality
ARQ	Automatic Repeat Request
AS	Access Stratum
AWGN	Additive White Gaussian Noise
BH	Backhaul
BLER	Block Error Rate
BPSK	Binary Phase Shift Keying
BS	Base Station
BSR	Buffer Status Report
CA	Carrier Aggregation
CAPEX	Capital Expenditure
CAC	Central Access Controller
CAC-C	Central Access Controller, Control Plane part
CAC-U	Central Access Controller, User Plane part
CDF	Cumulative Distribution Function
CP	Control Plane
CP-L/M/H	Control Plane – Low/Medium/High
CPRI	Common Public Radio Interface
CP-OFDM	Cyclic Prefix OFDM
CPF	Control-Plane Functions
CPU	Central Processing Unit

C-RAN	Centralized Radio Access Network
CRC	Cyclic Redundancy Check
CSI	Channel State Information
CU	Central Unit
CN	Core Network
D/A	Digital/Analog
DC	Dual Connectivity
DFT	Discrete Fourier Transform
DFT-s-OFDM	DFT spread OFDM
DL	Downlink
DRA	Dynamic Resource Allocation
D-RAN	Distributed Radio Access Network
DRX /DTX	Discontinuous Reception/Transmission
dPDCH	direct transmittable Physical Downlink Channel
DU	Distributed Unit
eLTE	enhanced Long-Term Evolution
eMBMS	Enhanced Multimedia Broadcast Multicast Service
eNB	evolved Node B
EoGRE	Ethernet over GRE
EPC	Evolved Packet Core
EPS	Evolved Packet System
Eth	Ethernet
EVM	Error Vector Magnitude
FBMC	Filter Bank Multicarrier
FEC	Forward Error Correction
FDD	Frequency Division Duplex
FDM	Frequency Division Multiplexing
FFS	For Further Study
FFT	Fast Fourier Transform
FIR	Finite Impulse Response
flops	floating point operations per second
FWA	Fixed Wireless Access

gNB	3GPP terminology for 5G base station
GTP – U/C	GPRS Tunnelling Protocol – UP/CP
GRE	Generic Routing Encapsulation
GW	Gateway
HARQ	Hybrid Automatic Repeat Request
IC	Interference Cancelation
ICI	Inter-Carrier Interference
ICIC	Inter Cell Interference Coordination
IDFT	Inverse DFT
IEEE	Institute of Electrical and Electronics Engineers
IETF	Internet Engineering Task Force
IFFT	Inverse FFT
IMT	International Mobile Telecommunications
INI	Inter-Numerology Interference
IoT	Internet of Things
IP	Internet Protocol
IPv4	Internet Protocol version 4
IPv6	Internet Protocol version 6
ITU	International Telecommunication Union
KPI	Key Performance Indicator
L2 / L3	Layer 2 / Layer 3
LCP	Logical Channel Prioritization
LoS	Line-of-Sight
LTE	Long-Term Evolution
LTE-A	Long-Term Evolution Advanced
MAC	Medium Access Control
MeNB	Master evolved Node B
MIMO	Multiple-Input Multiple-Output
ML	Maximum-Likelihood
MMSE-IRC	Minimum Mean Square Error Interference Rejection Combining
mMTC	Massive MTC
mmWave	Millimetre Wave
MTC	Machine Type Communications
MUX	Multiplexing

NACK	Negative ACK
NDI	New Data Indicator
NF	Network Function
NFV	Network Function Virtualization
NGMN	Next Generation Mobile Network (Alliance)
NIID	Network Interface Identifier
NR	New Radio
OAM	Operations, Administration & Maintenance
OFDM	Orthogonal Frequency Division Multiplexing
OPEX	Operational Expenditure
OQAM	Offset QAM
OOBE	Out-Of-Band Emissions
PA	Power Amplifier
PAPR	Peak-to-Average Power Ratio
PCell	Primary Cell
PDAP	Packet Data Association Protocol
PDCCH	Physical Downlink Control Channel
PDCP	Packet Data Convergence Protocol
PDN	Packet Data Network
PDU	Protocol Data Unit
P-GW	PDN Gateway
PHICH	Physical HARQ Indicator Channel
PHY	Physical
PMIP	Proxy Mobile IPv6
P-OFDM	Pulsed-OFDM
PPN	Polyphase Network
PRB	Physical Resource Block
PS	Packet Scheduling
PSAM	Pilot Symbol-Assisted Modulation
QAM	Quadrature Amplitude Modulation
QCI	QoS Class Indicator
QoS	Quality of Service
RACH	Random Access Channel
RAN	Radio Access Network
RAT	Radio Access Technology
RB	Resource Block

RF	Radio-Frequency
RLC	Radio Link Control
RN	Relay Node
RoHC	Robust Header Compression
RRC	Radio Resource Control
RRH	Remote Radio Head
RRM	Radio Resource Management
RS	Reference Signals
RTT	Round Trip Time
RU	Radio Unit
Rx	Reception
SAE	System Architecture Evolution
SAW	Stop-And-Wait
SCell	Secondary Cell
SC-FDMA	Single Carrier Frequency Division Multiple Access
SCS	Subcarrier Spacing
SDN	Software Defined Network
SDU	Service Data Unit
SeNB	Secondary evolved Node B
SFB	Synthesis Filter Bank
SFN	Single-Frequency Network
SINR	Signal to Interference and Noise Ratio
SNR	Signal to Noise Ratio
SPB	Shortest Path Bridging
s-gNB	Secondary gNB
S-TMSI	SAE Temporary Mobile Subscriber ID
TCP	Transmission Control Protocol
TDD	Time Division Duplex
TDM	Time Division Multiplexing
TeC	Technology Component
TEID	Tunnel Endpoint Identifier
TRILL	Transport Interconnection of Lots of Links

TWAG	Trusted Wi-Fi Access Gateway
TWAN	Trusted WLAN Access Network
TTI	Transmission Time Interval
Tx	Transmission
UCI	Uplink Control Information
UDN	Ultra-Dense Network
UDP	User Datagram Protocol
UE	User Equipment
UFMC	Universal Filtered Multicarrier
UL	Uplink
UM	Unacknowledged Mode
uMTC	Ultra-Reliable MTC
UMTS	Universal Mobile Telecommunications System
UP	User Plane
UPF	User-Plane Functions
URLLC	Ultra-Reliable Low Latency Communications
USTM	Unitary Space-Time Modulation
V2I	Vehicular-to-Infrastructure
V2V	Vehicular-to-Vehicular
V2X	Vehicular-to-X / Vehicle-to-Everything
VLAN-ID	Virtual Local Area Network-Identifier
VR	Virtual Reality
WiFi	Trademark of the Wi-Fi Alliance for WLAN technologies
WiMAX	Worldwide Interoperability for Microwave Access
W-OFDM	Windowed-OFDM
WRC	World Radiocommunication Conferences
x-haul	Cross haul
xMBB	Extreme Mobile Broadband
ZT-DFT-s-OFDM	Zero Tail DFT-s-OFDM

1 Introduction

1.1 Objective of the document

In comparison to the legacy communication systems, 5G systems have to be more flexible and scalable to support new heterogeneous use cases and devices. An essential question related to the 5G design is how the different air interface (AI) candidate technologies, including LTE-Advanced (LTE-A) evolution and those introduced in [MET216-D41], can be flexibly integrated into one overall 5G AI. Such design should maximally leverage availability of a wide portfolio of spectrum bands, radio channel characteristics, connectivity options, cell types etc. to address various new use cases and requirements, in a way that both the complexity of the standard and that of the implementation are minimized, while the performance of individual technologies is not sacrificed. Apart from the physical layer aspects, this includes higher radio protocol stack mechanisms and user plane functions in particular. This topic was addressed within METIS-II by Work Package 4 (WP4) in its earlier deliverable [MET216-D41], and the present deliverable aims at addressing some of the final underlying unknowns.

This deliverable contributes to one of the key innovation pillars developed in METIS-II related to the holistic air interface harmonization framework and complements and extends the work presented in previous WP4 deliverable D4.1 “Draft air interface harmonization and user plane design” [MET216-D41]. It contains the final concepts on air interface harmonization and user plane design. In addition to the control plane design work done in Work Package 5 [MET217-D52] and Work Package 6 [MET217-D62], both deliverables will serve as a key input to the overall 5G RAN design in deliverable D2.4.

The main objectives of this deliverable are:

- to provide an analysis of physical layer enablers (new waveforms, enhanced spectrum up to 100GHz, different frame structure and numerologies, non-coherent reception) to be introduced in the 5G context. We focus on multi-service coexistence considerations and harmonization, giving examples of various options and underlying trade-offs. Preliminary technical work on these aspects was extensively considered in chapter two, three and four of D4.1;
- to expand the analysis of a corresponding flexible 5G user plane design and harmonization with new protocol stack functions and mechanisms required to support diverse service requirements of the 5G service families, i.e., xMBB, mMTC, and uMTC. Deliverable D4.1 included an initial study of harmonized user plane design and also protocol aggregation approaches and this deliverable complements those considerations;
- to present different user plane architecture concepts including software defined architectures and a study how various user plane design decisions related to the

functional split options, a control / user plane split and Network Slicing impact on the overall 5G architecture.

1.2 Structure of the document

Continuing the first phase of work presented in D4.1, an adaptable and flexible 5G AI design is further studied to address the support for different services types envisioned in the 5G ecosystem. Key METIS-II design principles are therefore recommended which enable this design.

The harmonized AI framework includes design characteristics, functions and parameters at each layer of the radio protocol stack, starting from the physical layer mechanisms suitable for 5G, which are analysed in Chapter 2. First, based on the formulated harmonized implementation framework, a multi-waveform implementation is proposed taking into account the complexity and latency trade-offs. Furthermore, the coexistence of the broad landscape of 5G frequency bands below and above 6 GHz is studied in the context of their suitability for addressing heterogeneous xMBB, mMTC and uMTC service requirements. Particular focus is given to different numerologies and frame structures usage in the same band. Finally, a comparison between coherent and non-coherent reception techniques is described including their harmonization proposal.

Chapter 3 discusses a service flexible 5G user plane design, focusing on the air interface protocol and functional aspects. This encloses the identification of major challenges in the legacy approaches, the analysis of new requirements, and the introduction of some new mechanisms to address them. Since multi-connectivity is an important aspect, leveraging the availability and benefits of multi-AI variants, new approaches are discussed by means of designing Hybrid / Automatic Repeat Request (HARQ/ARQ) tailored for new scenarios with different functional splits and service requirements. Finally, some considerations on the need for new 5G QoS model and the evaluation of the user plane protocol interface options between RAN and CN are provided.

The interaction between the user plane design in the protocol stack and a new 5G RAN architecture which supports a functional split between a central and a distributed unit is another key question addressed in Chapter 4. Initially, protocol split options are characterized with a comparison of the interface performance requirements between the units. In addition, the aggregation of AI variants in different layers of the protocol stack is analysed, to exploit multi-connectivity solutions by associating a 5G device to multiple cell groups, bands and services simultaneously taking into account the harmonization of user plane protocol functions. At the end, a user plane split from the control plane and its variants in the system architecture is studied and also some insight into dependency between network slicing requirements and the user plane is presented.

We finalize the document with Chapter 5 where we summarize the key findings and discuss future work.

2 PHY layer aspects that enable multi-service support

2.1 Introduction

One of the objectives of the METIS-II project is to facilitate the design of a unique 5G AI that supports a plethora of heterogeneous wireless services, characterized by both stringent and, most importantly, dissimilar/conflicting QoS requirements. However, this objective does not imply that every single device should implement all functionalities. That would depend on the services the devices are meant for. Still, a unique AI, which may integrate a number of AI variants (AIVs), based on the proposed harmonization paradigm, which was initially explained in [MET216-D41], is expected to simplify the design, as well as the implementation principles of multi-service devices, provided that certain re-configurability-enabling protocols are developed, destined to support parallel services and enable service-switching.

Focusing on the PHY layer, enabling configurability for multi-service support is a profound challenge, due to the envisioned occurrence of fast service-switching and concurrently active services. The need to multiplex wireless services is envisioned, for example, with respect to applications related to Virtual Reality (VR) or Augmented Reality (AR), where, extremely low end-to-end latency levels need to be achieved, together with immensely large data rates, with the aim of providing almost real-time, real-life user experience. Furthermore, assisted/remote surgery constitutes another illustrative example, where, exceptionally high spectral efficiency needs to be simultaneously achieved with latency values of a few milliseconds, as well as wide coverage/high reliability; such need is also encountered when considering safety or disaster-related scenarios. In METIS-II, a multitude of AIVs and frequencies is exploited, in order to efficiently realize such service multiplexing [MET216-D51].

Related to this, in this chapter, we select some PHY layer enablers that have been identified during the METIS-II project. These enablers are representative of main trends in 5G PHY layer design to support the strongly heterogeneous 5G services; they were developed either as part of METIS-II work or elsewhere and they are enhanced/analysed here using the METIS-II AI design and evaluation framework [MET216-D41]. These enablers are: new waveforms (cf. Section 2.2), exploitation of new spectrum in up to 100 GHz bands (cf. Section 2.3), co-existence of different numerologies and frame structures (cf. Section 2.4), and non-coherent reception techniques (cf. Section 2.5).

2.2 Multi-waveform harmonization

Modulation and waveforms are one of the key physical layer components that determine the system throughput, reliability, and complexity, therefore their design is critical in meeting the various requirements of 5G services and addressing the challenges imposed by extremely diverse use cases, deployment scenarios and service requirements in 5G systems. These requirements and challenges include:

- *High spectral efficiency*: to satisfy the throughput requirement in 5G, waveforms are expected to have high spectral efficiency and low out-of-band emission to eliminate the need for guard bands.
- *Low Peak-to-Average Power Ratio (PAPR)*: One of the common disadvantages of multicarrier waveforms is their high PAPR that affects mainly the battery consumption at the terminal side. Lower PAPR is needed for 5G waveform design especially when high power efficiency is desired.
- *Compatibility with Multiple-Input Multiple-Output (MIMO) technologies*: 5G systems are expected to deliver an efficient use of the spectrum by using MIMO technologies (such as massive MIMO). Hence, waveforms that can efficiently support MIMO are very desirable.
- *Support for sporadic access*: In order to reduce the power consumption, (especially for mMTC), low-power devices require transmitting their data immediately after waking up with very low overhead and enter a dormant state directly after data transmission. Thus, as devices cannot be fully synchronized in this scenario, 5G waveforms need to be robust against timing and frequency offset to limit the amount of required signalling.
- *Robustness against Radio-Frequency (RF) impairments*: waveforms should in general address challenges caused by RF impairments, e.g., phase noise, I/Q-imbalance, sampling jitter, sampling frequency offset, carrier frequency offset, Power Amplifier (PA) nonlinearity, etc. In particular, RF impairments become more severe in millimetre Wave (mmWave) bands. As a result, robustness against RF impairments needs to be enhanced in such frequency bands.
- In addition to the main challenges listed above, there are many other issues that need to be taken into consideration when designing 5G waveforms such as time localization. Time localization refers to the time confinement of the transmitted symbols. It is well-known that time and frequency localization cannot both be improved without limits. A signal well time localized has to be expanded in frequency, and vice-versa. Time localization is important for new services that require short switching times in Time Division Duplex (TDD) systems, or low overhead for short packets transmission.

As mentioned in [MET216-D41], METIS-II has selected a comprehensive set of AIVs which help meeting one or more 5G KPIs, and conform to one or more 5G AI design principles. These are categorized into two main waveform families: Orthogonal Frequency-Division Multiplexing (OFDM) - based waveforms and Filter Bank Multicarrier (FBMC) - based waveforms.

2.2.1 3GPP waveforms study work

In 3GPP, waveforms for NR have been discussed, and for phase I and with focus on xMBB, some agreements have been achieved including the requirement to consider different waveforms and their harmonization. It has been agreed during RAN#86 meeting in August 2016 in Gothenburg that at least for frequency bands below 40 GHz for xMBB and URLLC services (similar to uMTC), NR supports Cyclic Prefix OFDM (CP-OFDM) for Downlink (DL) and Uplink (UL), possibly with additional low PAPR technique(s). NR should also support Discrete Fourier Transform spread OFDM (DFT-s-OFDM), at least for xMBB uplink for up to 40GHz. CP-OFDM can be used for single-stream and multi-stream (i.e. MIMO) transmissions, while DFT-s-OFDM is limited to single-stream transmissions. The network decides which waveform to use, but both CP-OFDM and DFT-s-OFDM are mandatory for User Equipments (UEs). It should be noted that from 3GPP Radio Layer 1 Working Group (RAN1) perspective, spectral confinement technique(s) (e.g. filtering, windowing, etc.) can be used for a waveform at the transmitter but they should be transparent to the receiver. It has been agreed that RAN1 should target a common framework for CP-OFDM and DFT-s-OFDM (without compromising CP-OFDM performance and complexity), i.e., control channels, Reference Signals (RS), etc. For DFT-s-OFDM, it is also agreed that NR should support 0.5*pi Binary Phase Shift Keying (BPSK) modulation. However, the details of frequency domain spectrum shaping will be further studied and the case where no spectrum shaping is needed is not precluded. For CP-OFDM, the ratio of transmission bandwidth configuration over channel bandwidth, Y, should be greater than that of LTE, i.e., Y>90%. Some evaluations in RAN1 show that Y can be up to 98% [3GPP16-166093].

Even though the current focus of waveform is on CP-OFDM and DFT-s-OFDM in phase I for below 40 GHz, NR is still open to other waveform options for above 40 GHz as well as for other services (e.g. mMTC).

2.2.2 Background to multi-waveform usage

CP-OFDM is the underlying waveform adopted by different wireless standards such as WiFi (Institute of Electrical and Electronics Engineers (IEEE) 802.11x trademark term), Worldwide Interoperability for Microwave Access (WiMAX) and LTE. The advantages of CP-OFDM include low baseband complexity, simple implementation to combat severe multipath fading, good affinity with MIMO systems, good time localization, and flexibility in numerology and frame structure design. On the other hand, classical CP-OFDM has comparatively high Out-Of-Band Emissions (OOBE), and additional signalling overhead due to the use of the cyclic prefix. It cannot handle the asynchronous or high mobility users well because of the Inter-Carrier Interference (ICI). Moreover, as aforementioned, high PAPR is an issue for any multicarrier waveform.

On the contrary, FBMC represents a multicarrier system where the single subcarrier signals are individually pulse shaped with a prototype pulse to achieve good spectral containment in the frequency domain, in such a way that sub-bands are isolated to allow for individual PHY

configurations. This fact renders FBMC an enabler for offering significantly enhanced degrees of freedom for the system design in the frequency domain. As aforementioned, sporadic traffic has emerged as an important service for 5G and achieving perfect synchronization in such scenario could be problematic. The good spectral containment leads to smaller ICI caused by Asynchronous transmission. FBMC also avoids using a cyclic prefix and therefore achieves higher spectral efficiency. However, the duration of one FBMC symbol lasts for multiple symbol intervals and therefore symbol overlapping will happen. Moreover, the baseband complexity will be higher than that of CP-OFDM systems due to the additional filtering as well as the modified baseband processing, as required for e.g. channel estimation.

Multi-waveform harmonization was not critical in LTE because the main target use case is xMBB, but in 5G, the use cases as well as the KPI requirements are much more diverse. So far, some experimental testbeds including multiple waveforms have been developed [DDF14, GVM+16, JS15, KKD+15, NNB+14, NNB16], but mainly focused on dedicated implementations for isolated waveforms, without considering a versatile implementation able to generate different waveforms according to a harmonized framework with hardware reuse (the impact of hardware reuse in multi-connectivity will be discussed in Section 2.2.4). In contrast to previous approaches, we propose a harmonized waveform implementation able to reduce the overall complexity and memory usage with respect to multiple isolated waveform implementations. As a first step towards waveform harmonization, Gabor systems [FS98] are a useful mathematical means to provide a general framework for multicarrier systems, where different multicarrier waveforms can be represented by selecting the appropriate prototype filter, subcarrier spacing (SCS) and symbol spacing in time. A comprehensive survey on multicarrier waveforms based on Gabor systems can be found in [SGA14] and references therein.

2.2.3 Multi-waveform implementation proposal

2.2.3.1 General framework

2.2.3.1.1 Transmitted signal

A multicarrier scheme is based on multiple subcarriers transmitted at the same time. Mathematically, the transmitted signal $x(t)$ is expressed as

$$x(t) = \sum_{m=-\infty}^{\infty} \sum_{k=0}^{N-1} X_{mk} g_{mk}(t), \quad (1)$$

where m is the time index (the number of multicarrier symbols in the time dimension is assumed to be infinite), k is the subcarrier index, X_{mk} is the modulated symbol transmitted at the k-th subcarrier and m-th multicarrier symbol (generally drawn from a set of complex numbers, although real numbers can be also considered as a special case), N is the number of subcarriers with frequency spacing $\nu = 1/\tau$, where τ denotes the symbol duration, and g_{mk} is

the pulse shape for the m-th multicarrier symbol and k-th subcarrier, which maps the symbols X_{mk} to the actually transmitted signal. The general multicarrier scheme is defined as a *Gabor system* [SGA14] when g_{mk} has the form

$$g_{mk}(t) = p(t - mT)e^{j2\pi k(t-mT)/\tau}, \tag{2}$$

where $p(t - mT)$ is called the transmitter prototype filter (or Gabor atom), and T is the multicarrier symbol interval.

Focusing on practical digital systems, the discrete-time version of the Gabor system describes the transmitted signal, denoted by $x[n]$, which can be obtained by sampling with period T_s

$$\begin{aligned} x[n] = x(nT_s) &= \sum_{m=-\infty}^{\infty} \sum_{k=0}^{N-1} X_{mk} p[n - mN_{\mathrm{T}}] e^{\frac{j2\pi kT_s}{\tau}(n-mN_{\mathrm{T}})} \\ &= \sum_{m=-\infty}^{\infty} \sum_{k=0}^{N-1} X_{mk} p[n - mN_{\mathrm{T}}] e^{\frac{j2\pi k}{N}(n-mN_{\mathrm{T}})}, \end{aligned} \tag{3}$$

where $N_{\mathrm{T}} = T/T_s$ is the length of the total symbol interval in samples and the symbol duration has been expressed in terms of T_s as $\tau = NT_s$. The discrete-time version of the prototype filter $p[n] = p(nT_s)$ has length $L = KN_{\mathrm{T}}$, with K standing for the overlapping factor.

Renaming $i = n - mN_{\mathrm{T}}$, the prototype filter $p[i]$ has the following form:

$$p[i] = \begin{cases} p_i \in \mathbb{C}, & \text{if } 0 \le i \le KN_{\mathrm{T}} - 1, \\ 0, & \text{otherwise.} \end{cases} \tag{4}$$

Note that the waveform exhibits overlapping in time between consecutive symbols because, from the prototype filter definition, $p[n - mN_{\mathrm{T}}] \neq 0$ when $(m+1)N_{\mathrm{T}} \le n \le (m+K)N_{\mathrm{T}} - 1$, and thus, in this context, K is also called the overlapping factor.

2.2.3.1.2 Received signal

Reconstructions of the modulated symbols originally transmitted from the received signal $x[n]$, denoted as $\tilde{X}_{mk}$, may be achieved by reverting the operations performed at the transmitter as follows

$$\tilde{X}_{mk} = \sum_{n=-\infty}^{\infty} x[n] e^{-j2\pi k(n-mN_{\mathrm{T}})T_s/\tau} p^*[-(n - mN_{\mathrm{T}})] + w[n], \tag{5}$$

with $w[n]$ being statistically uncorrelated samples of an Additive White Gaussian Noise (AWGN). Note that the signal detection operation involves convolving the received signal with a conjugated time-reversed version of the pulse shape and, thus, it is equivalent to matched filtering.

2.2.3.1.3 Waveform generation within the framework

The general framework presented above can be particularized to obtain a set of multicarrier waveforms under consideration for 5G. In particular, the mathematical framework is useful to represent several CP-OFDM variants, such as classical CP-OFDM, Windowed-OFDM (W-OFDM) [ABS11], Pulse-shaped-OFDM (P-OFDM) [ZSW+15] and Single Carrier Frequency Division Multiple Access (SC-FDMA) or Zero Tail DFT-spread-OFDM (ZT-DFTs-OFDM) [BTS+14]. Furthermore, the framework is also valid to represent waveforms of the FBMC family such as FBMC-Quadrature Amplitude Modulation (FBMC-QAM) [KYK+16] and FBMC-Offset QAM (FBMC-OQAM) [SSL02].

On the other hand, the Universal Filtered Multicarrier (UFMC) waveform performs a particular filtering per sub-band with filters that cannot be implemented based on Polyphase Network (PPN)-based filter banks. Therefore, this waveform does not directly fit within the presented framework, which is based on PPN filtering. In fact, as shown in theory, UFMC could neither be represented by a single particularization of Eq. (3) since UFMC structure goes beyond Gabor systems formulation and would roughly need as many Gabor systems as sub-bands to be represented and generated.

As a result, the proposed harmonized implementation will be focused on the set of waveforms directly matching the framework. Nevertheless, implementation considerations and the complexity analysis of UFMC will be also given in Section 2.2.3.5.

2.2.3.2 Implementation aspects of the harmonized transceiver design

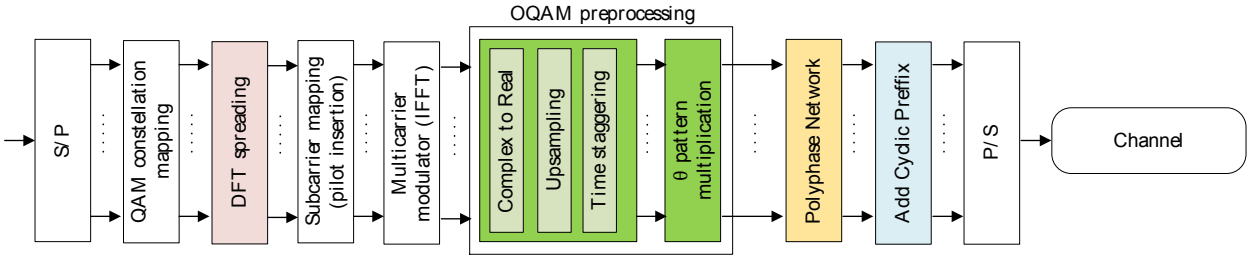


Figure 2-1: Proposed harmonized transmitter for multicarrier waveform generation.

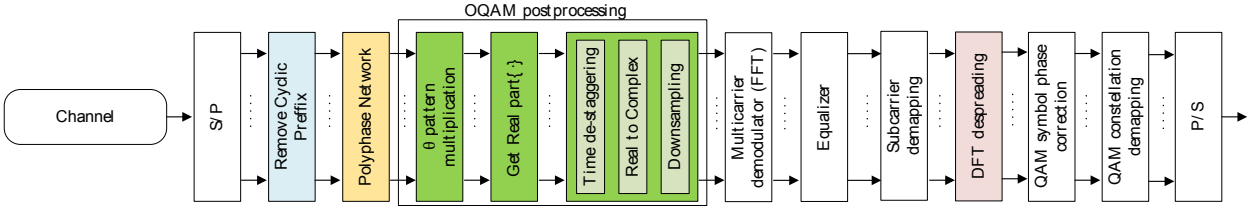


Figure 2-2: Proposed harmonized multicarrier receiver.

An all-encompassing solution integrating multiple waveforms in a single harmonized implementation would be very beneficial to provide flexible adaptation to a particular

communication scenario and, at the same time, to reduce implementation costs. For this particular purpose, the block diagrams of a proposed harmonized transmitter and receiver are presented in Figure 2-1 and Figure 2-2, respectively. The diagrams correspond to a generic multicarrier waveform as the one represented by Eq. (3), focusing on a specific implementation based on polyphase filtering, carried out in the time domain through a PPN [SSL02]. As further elaborated next, by selectively enabling or disabling particular blocks, both transmitter and receiver harmonized implementations are able to generate and reconstruct each one of the different multicarrier waveforms.

On the one hand, all the waveforms include all the general blocks of the diagram (in white). On the other hand, the blocks of the harmonized implementation that require specific configuration for each waveform are emphasized with colours and are next detailed:

- *Discrete Fourier Transform (DFT) spreading/despreading*: The pink blocks are intended to perform the spreading/despreading operations necessary for the ZT-DFT-s-OFDM waveform transmission/reception.
- *OQAM preprocessing/postprocessing*: The green blocks contain the necessary preparative multiplexing steps for the FBMC-OQAM transmission (complex to real number conversion of QAM complex symbols, upsampling, and time staggering), and FBMC-OQAM reception (real to complex number conversion, downsampling, and time de-staggering).
- *PPN*: The yellow blocks perform the convolution of the discrete signals with a filter implemented through a PPN. Throughout this deliverable we consider that the prototype filter used with the PPN has real coefficients, i.e., it is symmetric in the frequency domain.

The particularization of the harmonized block diagram to each multicarrier waveform illustrates the usefulness of the proposed implementation. Regarding the CP-OFDM variants, all of them will leave aside the green blocks and include the blocks in blue for the cyclic prefix addition and removal, except from ZT-DFT-s-OFDM that does not use cyclic prefix. With respect to the inclusion of the filtering blocks (in yellow), it will actually depend on the specific variant. Plain CP-OFDM will not require these blocks, whereas W-OFDM and P-OFDM will need them. ZT-DFT-s-OFDM will require the pink blocks. Concerning the FBMC-QAM waveform, it must contain all the blocks in Figure 2-1 and Figure 2-2 except for those involving the cyclic prefix, which are shown as blocks in blue, and the operations in charge of OQAM generation, shown in green. Finally, the transmission of FBMC-OQAM will require all the blocks of the diagram but the ones involving the cyclic prefix.

Further implementation aspects of particular transceiver blocks are discussed in Annex A.

2.2.3.3 Multi-waveform implementation complexity

In this section, we analyse the complexity of the multi-waveform implementation for which some of the complexity computations performed in Annexes A and B have been used.

2.2.3.3.1 Complexity of proposed harmonized transmitter

The minimum complexity cost of the harmonized implementation requires: one Fast Fourier Transform (FFT) block of size N (common to all waveforms) with complexity $C\mathrm{f}_{\mathrm{FFT}}(N)$ (cf. Annex A.1); one block to rebuild the FFT of two real-valued inputs (required for FBMC-OQAM) with complexity $C\mathrm{f}_{\mathrm{C2R}}(N)$ (cf. Annex A.1); two blocks of real-valued PPN filtering with complexity $C\mathrm{f}_{\mathrm{PPN\text{-}\Re}}(N)$ each (cf. AnnexA.3); and one block for DFT spreading of size M necessary for ZT-DFT-s-OFDM (cf. Annex A.2). The aggregated complexity cost is

$$C\mathrm{m}_{\mathrm{HARM}}(N,K) = N\log_2 N + 4NK + 5N + M\log_2 M - 3M + 8\,, \tag{6}$$

$$C\mathrm{a}_{\mathrm{HARM}}(N,K) = 3N\log_2 N + 4NK - 3N + 3M\log_2 M - 3M + 8, \tag{7}$$

$$C\mathrm{f}_{\mathrm{HARM}}(N,K) = 4N\log_2 N + 8KN + 2N + 4M\log_2 M - 6M + 16. \tag{8}$$

2.2.3.3.2 Complexity of non-harmonized transmitter

The complexity cost of a solution where all the waveforms are drawn together as standalone implementations is found by adding the number of multiplications, additions, or floating point operations per second (flops) in Eq. (A11)-(A28), and (A32)-(A34) in Annex B. The total result is

$$C\mathrm{m}_{\mathrm{NOHARM}}(N,K) = 6N\log_2 N + 8NK - 8N + M\log_2 M - 3M + 28, \tag{9}$$

$$C\mathrm{a}_{\mathrm{NOHARM}}(N,K) = 18N\log_2 N + 8NK - 22N + 3M\log_2 M - 3M + 28, \tag{10}$$

$$C\mathrm{f}_{\mathrm{NOHARM}}(N,K) = 24N\log_2 N + 16NK - 30N + 4M\log_2 M - 6M + 56. \tag{11}$$

2.2.3.3.3 Complexity comparison

Assuming the typical values $K = 4$ and $M = 12$, Figure 2-3 and Figure 2-4 show the complexity in terms of multiplications, additions and flops of the harmonized and non-harmonized implementations for different values of N. From basic calculations from the values in the figures, it can be observed that the typical savings range between 60–75%. Figure 2-5 shows the increase in the number of flops of the harmonized implementation with respect to the implementation of a single specific waveform. As shown, while the harmonized implementation is e.g. between 50–80% more complex in terms of flops than the CP-OFDM, ZT-DFT-s-OFDM and W-OFDM waveforms, the harmonized implementation cost increase tends to be negligible with N for the most complex waveform, i.e., FBMC-OQAM. In other words, if FBMC-OQAM is implemented, it is almost free (in terms of computational cost) to implement the other 5 waveforms in a harmonized way.

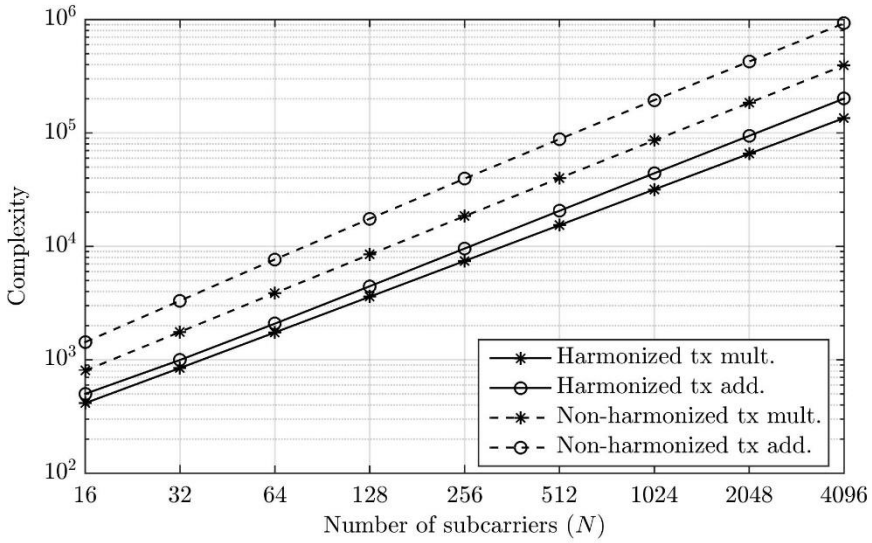


Figure 2-3: Multiplications and additions of the proposed harmonized and the non-harmonized implementation of the six waveforms.

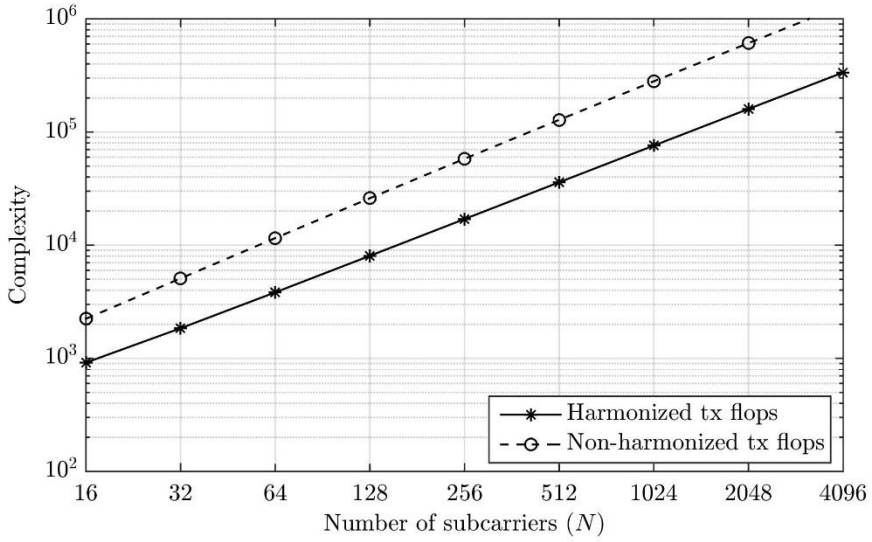


Figure 2-4: Flops of the proposed harmonized and the non-harmonized implementation of the six waveforms.

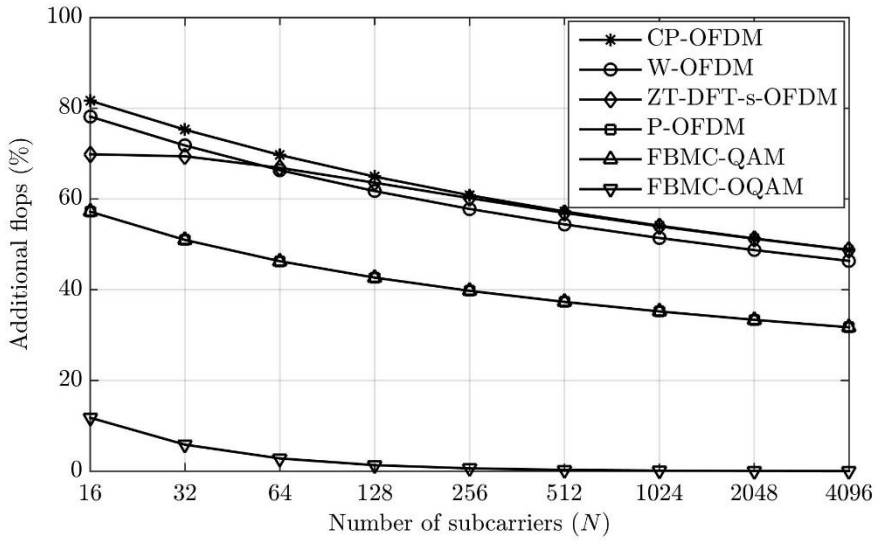


Figure 2-5: Additional flops needed by the harmonized implementation with respect to the implementation of a single specific waveform.

2.2.3.4 Chip space aspects

Figure 2-6: Covered chip space when the non-harmonized multi-waveform implementation approach is considered.

The considered individual waveforms are illustrated in Figure 2-6, where processing blocks of waveforms belonging to the OFDM and FBMC waveform families appear in blue and green colour, respectively. It can be observed that, for implementing a non-harmonized multi-waveform transceiver, thirteen processing blocks are needed in total: six FFT blocks, denoted as “FFT”, four blocks of real-valued PPN filtering where $K \geq 1$, denoted as "PPN$_R$(N, K)”, one block of real-valued PPN filtering with $K = 1$, i.e., “PPN$_R$(N, 1)”, as well as, a block for OQAM pre-processing, denoted as “OQAM pre-processing”, and a block for DFT spreading, denoted as “DFT spread”.

Focusing now on the harmonized approach, and following the analysis undertaken in Sections 2.2.3.2 and 2.2.3.3.1, merely five blocks are needed in order to realize a harmonized multi-waveform implementation, which are shown in Figure 2-7. As one would observe, a direct benefit of such an implementation option is that the required chip space is going to be compressed (i.e., only five implementation blocks are needed in total).

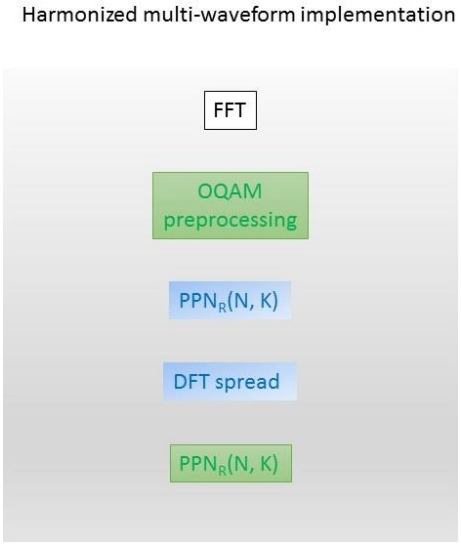


Figure 2-7: Covered chip space when the harmonized multi-waveform implementation approach is considered.

2.2.3.5 Implementation considerations of UFMC

As already mentioned in Section 2.2.3.1.3, the UFMC waveform requires an alternative implementation of the filter bank together with sub-band filtering. Therefore, this waveform cannot be included in the general framework and, thus, in the harmonized implementation proposal. In this section, the complexity of the UFMC waveform will be derived together with some additional implementation considerations that would affect the latency.

2.2.3.5.1 Alternative implementation of filter bank

Alternatively to PPN-based filtering, the filtering operation may be performed through the "overlap-add" method which zero-pads the input sequence. An efficient hardware implementation of zero padding consists of doubling the size of the FFT, so the cost is that of 2 FFTs of length $2N$ and N complex multiplications and $N-1$ complex additions for each of the $N+1$ samples of the filter, resulting in:

$$\mathcal{C}\mathrm{m}_{\mathrm{filter}} = 4N\log_2(N) - 4(N-2) \quad (12)$$

$$\mathcal{C}\mathrm{a}_{\mathrm{filter}} = 12N\log_2(N) + 2(N^2+N+3). \quad (13)$$

2.2.3.5.2 Complexity of UFMC

In the UFMC implementation considered, the B subbands have N/B subcarriers. At the transmitter, the subcarriers comprising the subband are spread over the full symbol length (an Inverse DFT (IDFT) spreader of size N must be performed for each subband) and then filtered before adding all of the sub-bands together and converting to the time domain. The total cost of the UFMC transmitter is

$$C\mathrm{m}_{\mathrm{TX/UFMC}}(N) = (3BN + 2N)\log_2 N + B(5N + 8) - 4N + 4 \quad (14)$$

$$C\mathrm{a}_{\mathrm{TX/UFMC}}(N) = (9BN + 6N)\log_2 N + B(N + 8) - 4N + 4 \quad (15)$$

$$C\mathrm{f}_{\mathrm{TX/UFMC}}(N) = (12BN + 8N)\log_2 N + B(6N + 16) - 8N + 8. \quad (16)$$

2.2.4 Investigating the latency performance of in-band multi-waveform harmonization

In principle, there exists a complexity-latency trade-off when considering harmonized and non-harmonized implementations. The harmonized implementation reuses some of the blocks for implementing the various isolated waveforms, in particular the FFT, which results in decreased complexity and chip size, as explained before. Although more complex and requiring more chip space, the non-harmonized implementation can, however, generate the multiple waveforms in parallel. Also, in the non-harmonized case, the computational latency of implementing the multiple waveforms is bounded by the processing time needed to implement the waveform of the highest complexity. Here, by computational latency, we refer to Central Processing Unit (CPU) clock cycles. The latency in seconds will depend on the CPU clock speed. Since waveforms need to be generated with some strict delays in milliseconds, an implementation with higher computational latency will require faster CPUs. Therefore, since complexity is also related to chip space, the complexity-latency trade-off can be understood as a chip space-clock speed trade-off.

Following the harmonized implementation framework and focusing on a worst-case scenario in which no hardware optimizations are considered and waveforms have to be generated in series, the resulting computational latency for multi-waveform multiplexing by means of harmonization is equal to the sum of the computational latencies of the individual waveforms. Fortunately, *optimized* harmonized implementations can also be realized which will help to achieve a better complexity-latency trade-off. Two examples are discussed next regarding multiplexing of two OFDM-based waveforms.

Efficient multiplexing of OFDM-based waveforms can be realized by using harmonized functional blocks, e.g. multiplexing of two ZT-DFTs-OFDM signals transmitted in different sub-bands as shown below.

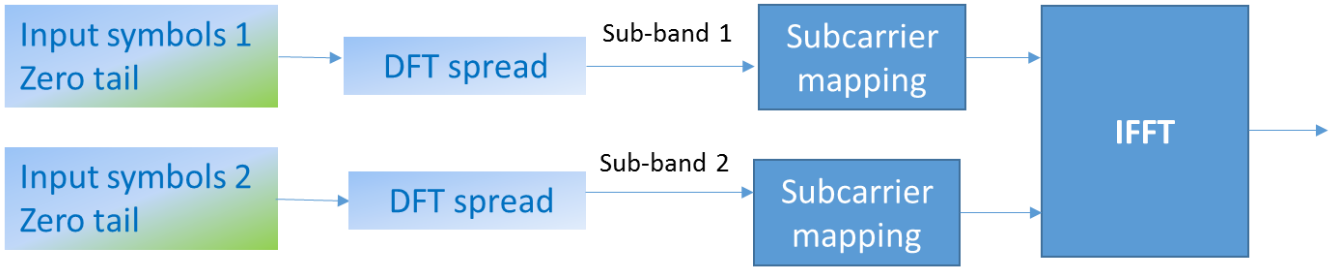


Figure 2-8: Multiplexing of ZT-DFTs-OFDM signals

For realization of the multiplexed waveform, the DFT spreading blocks can be executed either in parallel (two blocks have to be implemented) or in series (only one block is required). In parallel implementation computational complexity does not necessarily translate to latency. However in the series implementation, computational complexity effectively translates to computational latency. Section 2.2.3 provides computational complexity per sub-band, which is extended to the case of multiple sub-bands here. In particular, for a fixed DFT size M, and B number of sub-bands, the computational complexity (and hence the computational latency in series implementation) will be $O(M\log_2 M) \times B$. It is noted that the number of sub-bands $B = N/M$, and hence the computational latency performance has a behaviour of $O(M\log_2 M) \times N/M$. Moreover, ZT-DFT signal multiplexing advantageously uses only one IFFT block for multiplexing the sub-band signals with single numerology.

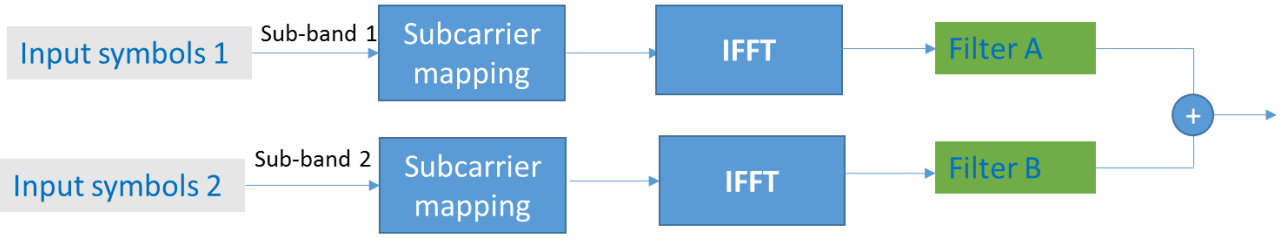


Figure 2-9. UFMC as multiplexing of filtered OFDM signals

As another illustration of hardware complexity vs computational time of multiplexing waveforms, Figure 2-9 shows multiplexed transmission of filtered OFDM signals corresponding to different sub-bands. In this case, multiple Inverse FFT (IFFT) blocks are realised for multiplexing the sub-band signals. In addition, the filtering blocks can be executed in parallel or in series based on the desired trade-off.

In the case of series implementation, the IFFT output of one sub-band will be applied by a filter A and then another IFFT output will be applied by a filter B serially in time using the same filter block. The filtered time domain signals are then superimposed to generate a multiplexed waveform. The computational latency is now a result of applying the filtering in series, i.e., as compared to a stand-alone implementation in parallel the latency is twice because of using the same filtering functional block in series for saving the hardware complexity. Thus OFDM

variants can realize a desired hardware complexity vs computational time trade-off by selectively reusing some of the functional blocks.

In the case of multiple numerology, the IFFT operation is expected to be applied separately per numerology. Thus one could expect that multiple IFFT operations need to be realized for multiplexing OFDM signals with different numerologies.

2.3 Considerations about the co-existence of multiple bands

The availability of multiple bands (above and below 6GHz) for 5G is expected to characterize 5G deployments. Due to the unprecedented traffic demand of next xMBB services, 5G will integrate mmWave bands with large available spectrum bandwidth. In this regard, the use of mmWave band is considered as one of the most promising approaches to significantly boost the capacity.

The spectrum landscape for 5G can be roughly split into the following three ranges of bands: low frequency bands below 6 GHz, middle frequency bands below 30 GHz, and high frequency bands above 30 GHz. It is expected that the channel conditions tend to be harsh as the frequency increases. In particular, the path loss attenuation, non-line of sight path loss exponent and penetration losses typically tend to increase with frequency, as has been reported in [Sam15] and International Telecommunication Union (ITU) feasibility study report [ITU13]. On the other hand, higher frequency bands also potentially offer wider bandwidths. Basically, from carrier frequency perspective, lower bands are expected to be used for coverage purposes and are therefore suitable for both Machine Type Communications (MTC) and outdoor xMBB services within large cell sizes where the terminal speed is expected to range from medium to high. On the contrary, higher bands with large contiguous bandwidth are more suitable for traffic of xMBB services for indoor and outdoor in small cells and with low terminal speeds.

Spectrum framework for 5G system is critical as it needs to flexibly balance the trade-off between providing enough system capacity and fulfilling different use cases requirements while taking into account increasing complexity of the radio interface and limitations of current spectrum regulation. This problem has been addressed in [MET216-D31]. Basically, it can be concluded that the suitability of spectrum bands for the three 5G service types are as follows (cf. Figure 2-10):

- For xMBB, a mixture of frequency spectra comprising lower bands for both coverage and high data traffic, and higher bands with large contiguous bandwidth to cope with the ever-increasing traffic demand, including wireless backhaul solutions, is required.
- For some mMTC applications, frequency spectrum below 6 GHz is more suitable and spectrum below 1 GHz is needed in particular when large coverage areas and good outdoor to indoor penetration are needed. Higher frequency bands are expected of less

relevance for this service type because of their small cell sizes and radio propagation challenges.

- For uMTC such as safety Vehicular-to-Vehicular (V2V) and Vehicular-to-X/Vehicle-to-Everything (V2X) communication, the frequency band below 6 GHz is an option. The sub-1GHz spectrum is particularly well-suited for high-speed applications and rural environments. The channel characteristics in mmWave bands can be challenging for this service type. However, one could also potentially use the mmWave frequency bands, especially the lower mmWave bands, by beneficially exploiting Line-of-Sight (LoS) propagation. In particular, short and frequent messages for vehicular safety could be interchanged with neighboring vehicles using mmWave. Thanks to the propagation conditions in this band, these messages would not propagate to distant vehicles, allowing closer vehicles to reuse the same band and, hence, increasing the capacity.

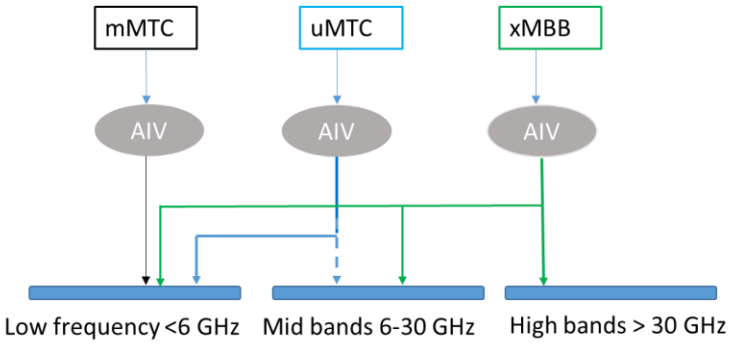


Figure 2-10: Mapping of service types to frequency ranges. The dotted line indicates a non-standalone operation in that band.

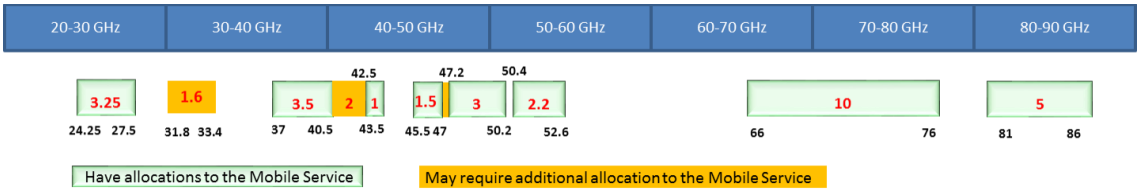


Figure 2-11: Frequency bands to be studied in ITU-R for IMT-2020 until WRC-19.

The spectrum allocation and availability for 5G above 6GHz has some limitations due to the existing regulatory framework. World Radiocommunication Conferences (WRCs) are held every three to four years and it is the job of WRC to review, and, if necessary, revise the Radio Regulations, the international treaty governing the use of the radio-frequency spectrum and the geostationary-satellite and non-geostationary-satellite orbits. WRC-15 agreed on a WRC-19 Agenda Item (1.13) to consider the identification of frequency bands for the future development of International Mobile Telecommunications (IMT), including possible additional allocations to the mobile service on a primary basis, in accordance with Resolution 238 (WRC-15). This involves conducting and completing the appropriate sharing and compatibility studies for a number of bands between 24-86 GHz in time for WRC-19 (cf. Figure 2-11).

In addition to the identified bands in WRC-15, some significant regional markets intend to continue to develop 5G in some other bands, e.g., 28 GHz in the US and Korea. A key aspect for 5G spectrum deliberations is to achieve global (or regional) harmonization to enable economy-of-scale benefits for 5G equipment. Sufficient harmonization does not only rely on having exactly the same spectrum available in different regions, which may be anyhow very difficult to achieve due to differing situations in different countries with regard to incumbent spectrum users and usage conditions. Optimized spectrum harmonization is also important from hardware implementation limitations, e.g. user device form factor, battery, transmit power, antenna limitations and cost perspectives. Combination of too many bands is prohibitive to be implemented in a device, it may fragment the device market unnecessarily and increase the device cost and size. In this context, harmonization can offer a solution for flexible global device implementation to maximize the targeted market for any device. In addition, some considerations should also be given to backwards compatibility with legacy systems, e.g. LTE, especially during the initial 5G deployment phase.

2.4 Considerations about the co-existence of different numerologies and frame structures in the same band

Although different numerologies and frame structures appear naturally in different bands, in case a unique band is available, the different service characteristics may require the co-existence of different numerologies and frame structures in this band. In this section, we will motivate the need of having different numerologies and frame structures in the same band, summarize 3GPP work on this topic and propose supporting Technology Component (TeC) solutions.

2.4.1 Different numerologies in the same band

Robust and efficient numerology and frame structure design are needed to achieve high throughput and reliable communication. Moreover, much lower latency compared with LTE/LTE-A should be supported with scalable design approaches. Besides considering the very diverse requirements imposed by different 5G services, a single numerology is unlikely to be optimized for all of them.

It has been agreed in 3GPP that forward compatibility of NR shall ensure smooth introduction of future services and features with no impact on the access of earlier services and UEs. In this regard, multiplexing different numerologies within a same NR carrier bandwidth (from the network perspective) should be supported, but different multiplexing options, such as Frequency Division Multiplexing (FDM) and/or Time Division Multiplexing (TDM) multiplexing can be

considered. NR allows flexible support of multiple numerologies to flexibly apply multiple SCSs (e.g. 15 KHz scaled by an integer) considering aspects of deployment scenarios (e.g. carrier frequency, cell size, Doppler spread, etc.) and target services (e.g. enhanced Multimedia Broadcast Multicast Service (eMBMS), V2X, URLLC, etc.) [3GPP16-RAN1-86]. Discussions on this topic have been ongoing within the 3GPP community, and several agreements and working assumptions have been reached. In the following, we will try to summarize the most relevant of these agreements and working assumptions [3GPP16-RAN1-86] [3GPP16-RAN1-86b].

- The LTE numerology with SCS of $f_0 = 15$ kHz is considered the reference numerology.
- Different SCSs for different numerologies should be restricted to a value of $2^n f_0$, with n being an integer.
- The number of subcarriers within one Resource Block (RB) is fixed to 12 for all numerologies.
- For numerologies with SCS of $2^n f_0$, the subcarriers are mapped on the subset/superset of subcarriers of SCS of f_0 in a nested manner in the frequency domain.
- The cyclic prefix overhead is assumed to be the same for all different numerologies (i.e. cyclic prefix duration is scaled with the different numerologies).

With all these listed agreements, a nested RB structure among different numerologies is achieved for FDM as shown in Figure 2-12.

Figure 2-12: Nested RBs structure for different numerologies.

The main enabler in NR for supporting multiple numerologies is the support of multiple SCSs, which can be beneficial for satisfying the requirements of diverse verticals and deployment scenarios. For example, for eMBMS, it would be natural to support Single-Frequency Network (SFN) transmissions to minimize coverage holes and provide more robust performance. For efficient SFN transmission, cyclic prefix has to be designed large enough to take into account the propagation delay of transmissions from multiple sites. Since larger cyclic prefixes lead to larger overhead, it inevitably requires longer OFDM symbol lengths than unicast transmissions and therefore smaller SCSs. On the other hand, larger SCSs are desirable for transmissions to or from high speed UEs. When a UE speed is very high (e.g. NR is targeting so support speeds up to 500km/h), the Doppler shift increases, and it is thus beneficial to configure larger SCS. Since NR is trying to consider a single framework supporting various use cases and deployment scenarios, multiplexing of different numerologies within a carrier bandwidth is becoming a favourable solution.

To support multiple numerologies within a carrier bandwidth, two approaches can be considered. One approach is FDM of multiple numerologies, which allows more flexibility in

scheduling UEs with different numerologies. However, when different numerologies are used in adjacent frequency resources at the same time, mutual Inter-Numerology Interference (INI) would be introduced between adjacent resources due to the violation of orthogonality between subcarriers of different numerology. In the following subsections, we will discuss several solutions to mitigate the impact of INI. The second approach is TDM of multiple numerologies. This approach has already been supported in LTE to multiplex eMBMS and unicast transmission. Compared to FDM, the operation of TDM is relatively simple. For TDM, there is no need to consider INI since only a single numerology exists per Transmission Time Interval (TTI). Therefore, relatively simple filtering and/or windowing are required. Moreover, there is no need to consider guard tones between frequency resources. On the other hand, TDM would restrict scheduling flexibility of UEs supporting different configured numerologies down to the time unit of a TTI, which may impact the latency in the system. In the rest of this section, we discuss some of the issues and challenges of FDM.

2.4.1.1 Spectral confinement techniques.

Multiplexing different numerologies in the same band will introduce INI due to the non-orthogonality of the adjacent frequency resources. To reduce the INI impact on the system performance, several spectral confinement techniques (e.g., filtering, windowing) have been discussed and evaluated [3GPP16-RAN1-86b] [3GPP16-167376]. Waveforms with low OOBE are more suitable to alleviate INI compared to the conventional CP-OFDM. In addition, guard tones can be inserted between frequency resources to facilitate fulfilling the strict OOBE requirement while simplifying filter design. A third approach is a transparent spectral confinement [BKW+17a]. This approach uses the fact that the receiver does not need to know exactly the filter impulse response or the window used at the transmitter side, but only some minimum requirements for the filter or window to be used (e.g., spectral and temporal masks, Adjacent Channel Leakage Ratio (ACLR), Error Vector Magnitude (EVM), etc.). These requirements specify only the features that are essential for the performance of the system. This gives flexibility for implementations, reduces signalling overhead between the transmitter and the receiver, and also opens opportunities of new advanced techniques in the future.

2.4.1.2 Resource grid – Guard band.

INI may become unacceptable if the interfering user has much higher power even when spectral confinement techniques are applied - e.g. in an uplink transmission where no proper power control could be applied e.g., in Internet of Things (IoT) applications due to the potentially relaxed synchronization or in downlink transmission with very large SCSs. In these situations, additional guard bands may be needed to achieve an acceptable level of performance.

The width of the required guard bands highly depends on the specific conditions, including modulation and coding schemes, power of neighbouring users and filter shape. In order to have

an efficient system, such guard bands should be chosen flexibly for any particular transmission. In addition, in many communication systems (e.g., in LTE) a regular resource grid is used, meaning that the entire resource space is divided into several RBs of the same size as shown in Figure 2-13 (a). The problem that may then occur is that, when introducing the guard bands of flexible size between the sub-bands, the sub-bands will be shifted apart in frequency domain and therefore violate the regular resource grid as illustrated in Figure 2-13 (b). The resulting irregular grid would lead to an increased overhead for signalling the shifted sub-band locations, since a simple index of regularly quantized RBs would no longer be sufficient to exchange channel status reports or scheduling decisions. To solve this problem, we propose allocating fractional RBs guard bands within the regular resource grid as illustrated Figure 2-13 (c) [WBK+16a] [WBK+16b] [MET216-D41].

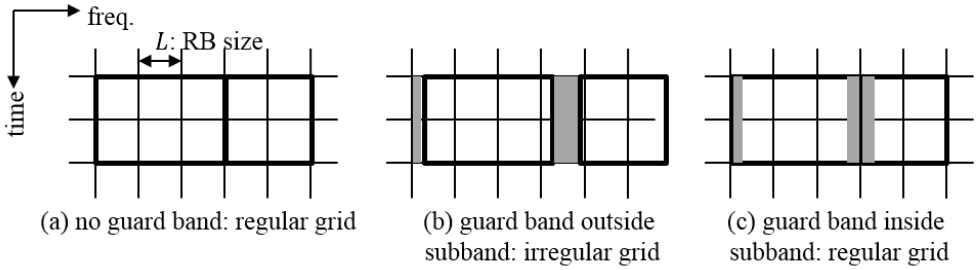


Figure 2-13: Guard band impact of the resource grid [BKW+17b].

2.4.1.3 Resource grid – Grid alignment.

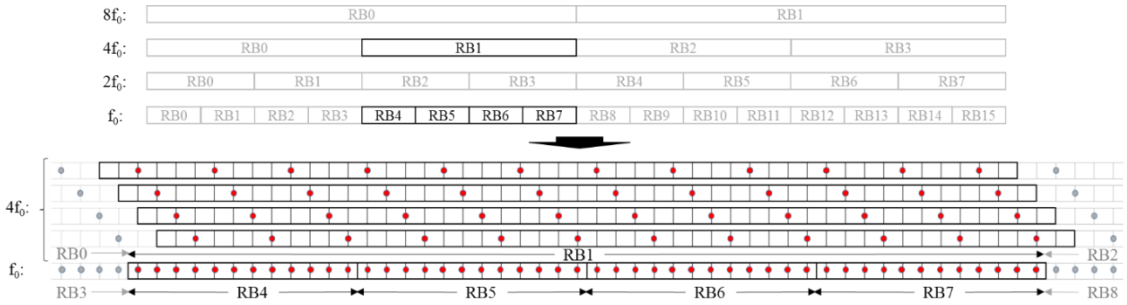


Figure 2-14: An example of 4 possible options/ambiguities of RB boundary for SCS of $4f_0$ without exceeding the RB grid boundaries of the reference numerology [BKW+17b].

Another resource grid aspect is the grid alignment among numerologies where several key agreements were made in 3GPP, as discussed earlier in this section. However, as we can observe in Figure 2-14 for example, there are still some possible ambiguities on how to locate the subcarrier centre frequencies (red dots) within each RB for different SCSs, or, in other words, how to define RBs boundaries while following the nested RB structure.

To understand the impact of the ambiguous options and how such ambiguities above could be resolved, we start with a straightforward approach illustrated in Figure 2-15, where we consider four different SCSs of f_0, $2f_0$, $4f_0$, and $8f_0$, just as an example. For all the different SCSs, the first subcarriers are aligned with the first subcarrier of the reference SCS f_0. In this approach, however, it is observed that interference to neighboring RBs is strongly biased to the left, in particular for wider SCSs. Another straightforward approach would be to have subcarriers equally and symmetrically spaced within the RB as illustrated in Figure 2-16, yielding an offset for larger SCSs at the edge of the RB. This approach, however, violates the aligned nested subcarrier grid and increases the complexity of the system due to the non-aligned FFT/IFFT sampling points for the different numerologies. Also the zero crossings of the subcarriers from different numerologies are not aligned anymore, which will result in increased interference.

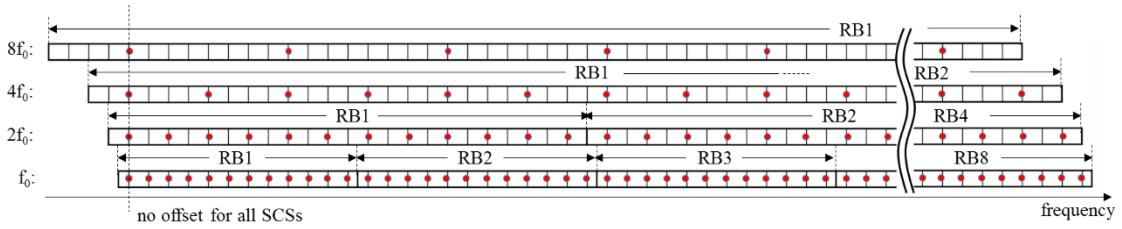


Figure 2-15: Straightforward approach 1: align the first subcarriers of all the numerologies [BKW+17b].

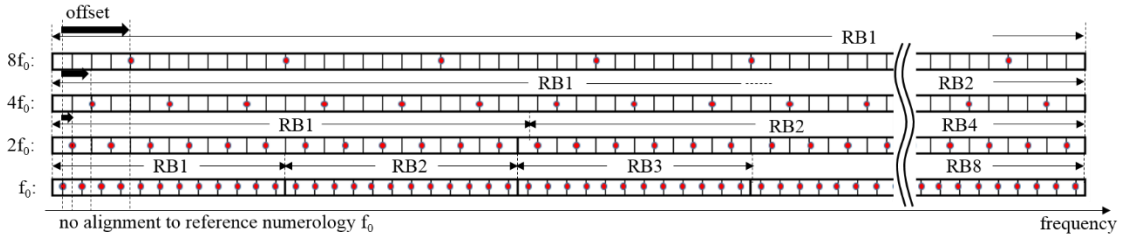


Figure 2-16: Straightforward approach 2: subcarriers equally and symmetrically spaced [BKW+17b].

To alleviate the strong interference bias while keeping the aligned nested subcarriers mapping, we propose introducing some offsets for centre subcarrier frequencies in Figure 2-15 according to the size of SCSs [BKW+17b] [3GPP16-1612724]. Such a frequency offset could be achieved by multiplying the time domain signal with a phase shift. To be more specific, there are 2^n choices as offset values (including zero offset) for SCSs of $2^n f_0$, without exceeding the RB grid boundaries based on the reference numerology. For example, for $n = 1$, i.e., SCS of $2f_0$, there are only two possible offset values; either no offset or offset f_0 to the right. Both options for $n = 1$ result in no essential difference; interference is slightly biased to the left or to the right. However, for $n > 1$ the proposed offset has more impact on the interference to neighbor RBs. For $n = 2$ (or $n = 3$), i.e., SCSs of $4f_0$ (or $8f_0$), there are 4 (or 8) choices for the offset values. The lower part of Figure 2-14 shows the 4 offset values for $n = 2$.

Figure 2-17 and Figure 2-18 illustrate two examples of introducing offsets while satisfying the nested subcarrier mapping rule. In both the figures, the interference to neighbouring RBs is more balanced as compared to Figure 2-15. In Example 1 in Figure 2-17, we introduce an offset of 30 kHz for SCSs of $4f_0$ and $8f_0$, while in Example 2 in Figure 2-18 we introduce an offset of 60 kHz for SCSs of $8f_0$. Offset values other than those in Examples 1 and 2 are also possible while keeping the nested subcarrier mapping rule. In the next section, we will evaluate these different approaches with link level simulations.

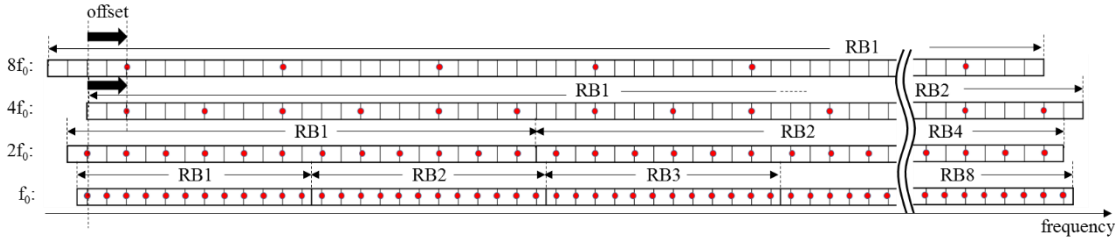


Figure 2-17: Example 1 of offsets to the first subcarrier for some SCSs, while keeping the nested subcarrier mapping [BKW+17b].

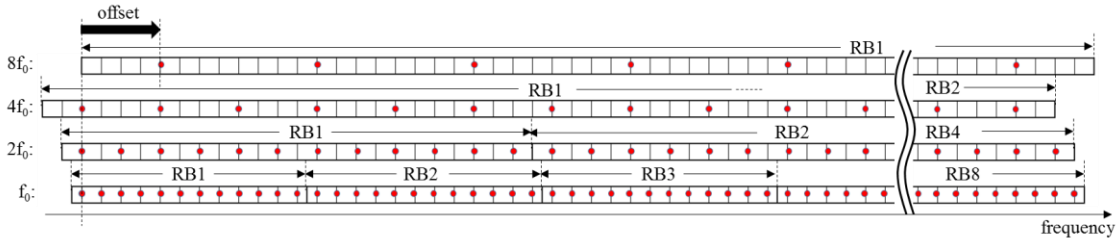


Figure 2-18: Example 2 of offsets to the first subcarrier for some SCSs, while keeping the nested subcarrier mapping [BKW+17b].

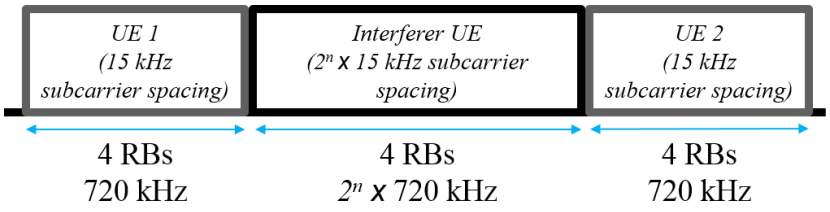


Figure 2-19: Evaluation setup [BKW+17b].

In the following, we compare the performance of the different grid alignment approaches under the setup shown in Figure 2-19. We assume a downlink setup with three users: two users of interest (termed as "target UEs") and one interfering UE located in the middle. The target UEs use a 15 kHz SCS, while the interfering user uses the $2^n 15$ kHz SCS. Each user is allocated 4 RBs, with one RB consisting of 12 subcarriers, i.e., the transmission bandwidth is 720 kHz for the target UEs and $2^n 720$ kHz for the interfering UE, respectively. We also assume zero guard

band between the UEs. We assume that the interfering user uses SCS of either 60 kHz ($n = 2$) or 120 kHz ($n = 3$). The assumed waveform is the conventional CP-OFDM.

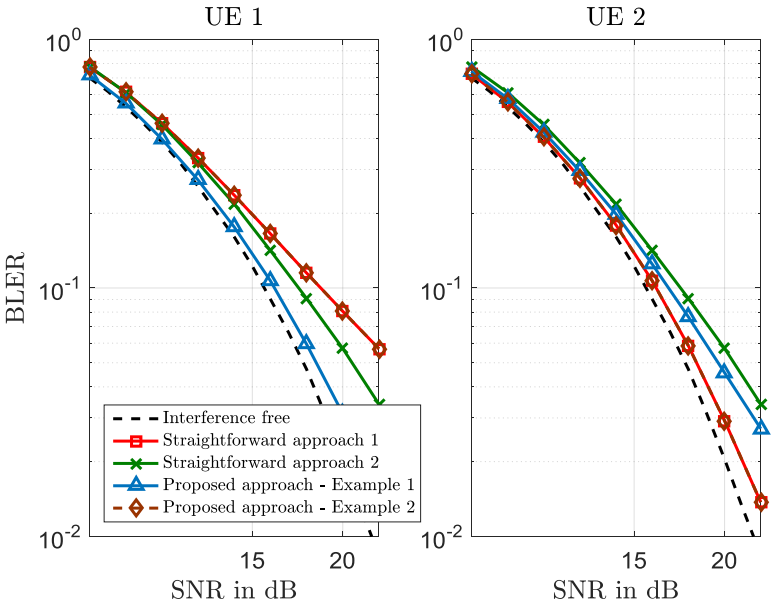


Figure 2-20: BLER vs. SNR results - 60 kHz SCS interfering user [BKW+17b].

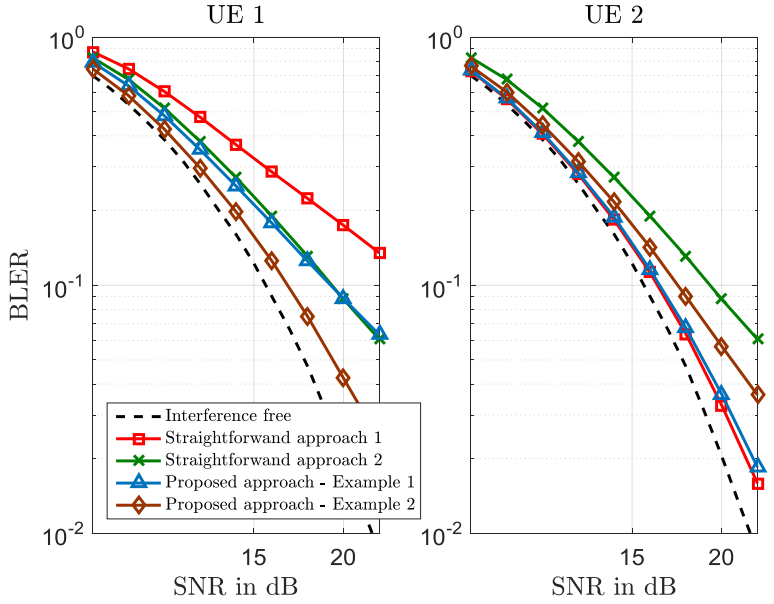


Figure 2-21: BLER vs. SNR results - 120 kHz SCS interfering user [BKW+17b].

Figure 2-20 and Figure 2-21 show the Block Error Rate (BLER) performance versus Signal to Noise Ratio (SNR) for the two target UEs with interfering UE using SCS of 60 kHz and 120 kHz, respectively. We could observe that UE1 suffers from the high interference bias in the case of

aligning the first subcarriers of all the SCSs (i.e., straightforward approach 1 shown in Figure 2-15). When the subcarriers are equally and symmetrically spaced within the RB (i.e., straightforward approach 2 in Figure 2-16), the interference becomes balanced between UE1 and UE2, but since the zero crossings of the subcarriers of different numerology are not aligned in this case, more subcarriers of the target UEs (with 15 kHz SCS) are interfered more severely, and hence cause the overall poor performance of both UEs.

The proposed approach (i.e., Examples 1 and 2) shows more balanced interference situations for both UE1 and UE2 while keeping the aligned nested subcarriers mapping. When we take a closer look into Figure 2-20 and Figure 2-21 and compare the performance of Example 1 (cf. Figure 2-17) and Example 2 (cf. Figure 2-18), we can make further observations as follows. The interference to neighbouring RBs caused by Example 1 in Figure 2-20 is more balanced than that in Figure 2-21, but more balanced interference is achieved with Example 2 in Figure 2-21. Considering that interference to neighbouring RBs is stronger with wider SCS (i.e. 120 kHz), the offset values of Example 2 may lead to less critical interference than those of Example 1. On the other hand, if 120 kHz is not often used or not needed at all, i.e., the offsets of Example 1 for up to 60 kHz lead to more balanced interference to neighbouring RBs than those of Example 2. Therefore, the proper choice of SCS-dependent frequency offset values depend on the range of SCSs to consider.

2.4.2 Different frame structures in the same band

For NR air interface, there is a need to define a minimum time and frequency structure, similarly to the Physical Resource Block (PRB) specified in LTE. A PRB consists of multiple subcarriers (N_{sc}) with SCS (Δf) occupying a total bandwidth of $B_{\mathrm{PRB}} = N_{\mathrm{sc}}\Delta f$, and has a duration in time of T_{PRB}. In LTE, although multiple transmission bandwidths from 1.4 MHz to 20 MHz are supported, constant value of SCS (e.g., $\Delta f = 15$ kHz) and fixed number of subcarriers in a PRB ($N_{\mathrm{sc}} = 12$) are used. The NR system should support a very wide range of spectrum, ranging at least up to 100 GHz, and much diverse transmission bandwidths.

Decision on the PRB size in NR could be affected by the minimum payload size to be supported in NR, since small PRB sizes (i.e., small size of $B_{\mathrm{PRB}} \times T_{\mathrm{PRB}}$) would be efficient in terms of resource utilization for small payload sizes, but inefficient in terms of L1/L2 signalling because more signalling is needed. It should be noted that the minimum payload size could be different depending on the applications and use cases in NR. Hence, supporting variable PRB sizes in NR could efficiently support various applications and use cases.

One of the key factors of frame structure is the slot duration. If we assume each slot consists of y symbols, we can choose $y = 7$ and $y = 14$ as initial settings for backward compatibility. The actual y value used in the system has to be signaled, hence increasing signaling overhead. However, the different quantities of symbols per slot to be considered are tightly related to a mini-slot structure. Mini-slots were recently proposed in 3GPP to provide smaller scheduling

units. If a mini-slot is defined as 7 symbols, then taking $y = 7$ is not needed. As another example, if the mini-slot is defined as 2 and/or 3 symbols and if mini-slot aggregation is supported, then taking $y = 7$ would be beneficial in terms of signaling overhead rather than $y = 14$. For example, with the assumption of 2 symbols in a mini-slot, $y = 14$ will have more options in an aggregated slot, e.g., as 2 up to 7 mini-slots can be aggregated within one slot. It will then cause more signalling overhead if all these options should be supported, compared to the case of $y = 7$, which only supports aggregation of 2 and 3 mini-slots.

However, signalling is not the only factor that influences slot duration, the numerology should be also taken into account. For 15 kHz SCS, a slot length of 14 symbols will not be able to meet xMBB user plane latency (4 ms) requirement in a TDD system considering the processing delay. Hence, using 7 symbols is preferable over using 14 symbols. For 30 kHz SCS, both $y = 7$ and $y = 14$ will be able to guarantee xMBB user plane latency. However, taking 7 symbols will cause more overhead in a dynamic TDD system, because it would require more frequent DL/UL switching within a subframe and this results in an increased number of gap symbols to be inserted between DL and UL symbols. So, it would be desirable to use 14 symbols as a slot for 30 kHz. For 60 kHz SCS, taking 14 symbols will give potential benefits in some particular scenarios as compared to using 7 symbols. For example, in a hybrid beamforming system, this will give an opportunity to facilitate rapid beam sweeping in a given slot. Moreover, similarly to 30 kHz SCS, a slot length of 14 symbols will allow for efficient utilization of the resources in a dynamic TDD system. From latency perspective, thanks to the short symbol duration with 60 kHz SCS, xMBB user plane latency requirement can be achieved even if we take a slot length of 14 symbols. In short, when different numerologies co-exist, co-existence of different frame structure should also be considered, enabling to optimize the signalling and latency performance.

Finally, another factor that influences frame structure in mmWave bands is beam-training. In particular, the use of very directional beams can severely affect the effective connection latency and the terminal connectivity itself, since the application of beam-training and beam-tracking protocols for Channel State Information (CSI) acquisition can create considerable overheads. To address the trade-off between reducing the resource (i.e., time, power) overhead during initial access and achieving satisfactory beamforming gains, several beam-training protocols have been proposed in current technical literature, i.e., in [DPG+16] [LJK+15] [PDG+16] [TPA11]. It should be noted, however, that the performance of these protocols depends on the available infrastructure (e.g., antenna elements, RF front-ends) both at the Access Point (AP) as well as at the device side. Consequently, to optimally exploit the limited degrees of freedom due to the application of hybrid (and not digital) beamforming, appropriate frame structure designs need to be considered. Assuming that the main source of communication latency is the application of the initial access protocol, this means that

- the design of a beam-training codebook comprising many extremely directional beams should lead to the reservation of considerable time resources for beam-search, since if only a subset of these beams are chosen for beam-training, the likelihood of the

occurrence of misalignment events will dramatically increase, hence seriously degrading the achieved UE throughput;

- when a few, wide "probing" beams are considered, for quickly establishing a connection with devices running delay-critical applications, the corresponding beam-training time can be shortened. This will occur, since the size of the beam-training codebook is not that much increased, as compared to the previous case, in other words, the whole 360-degree azimuth area can be "scanned" during only a few training stages.

Given the envisioned parallel existence of heterogeneous wireless services, and the capability of a mmWave AP to design more or less directional beams, balancing the available time resources of a frame towards establishing a connection of acceptable desired signal strength on one hand, and performing data transmission on the other hand, is a quite challenging problem to be tackled, especially from an interference mitigation point of view. The previously described throughput versus latency trade-off has been quantitatively investigated in [FDP+17], where, a standalone, single AP, multi-UE mmWave system is examined from a resource (i.e., time, hardware) optimization standpoint.

2.5 Harmonized coherent and non-coherent reception

2.5.1 Introduction and motivation

Most of the current communication systems are based on coherent reception where CSI should be estimated for equalization, demodulation, etc. The CSI can be obtained via training by inserting pilots in the data signals. This communication method is known as Pilot Symbol-Assisted Modulation (PSAM). These schemes, however, increase the signalling overhead, which can be counterproductive in systems with many antennas or fast movement of either side of the communication link [BH03] [MET216-D41].

This drawback increased the interest in schemes that do not require full CSI at the transmitter and the receiver. For the case of multi-antenna Rayleigh block-fading channels and motivated by [MH99], a communication method using Unitary Space-Time Modulation (USTM) and non-coherent reception was proposed in [HM00]. In this method, the signals transmitted from the antennas, viewed as matrices with $T \times M$ dimensions, where $T \geq M + N$ is the channel block length in channel uses, and $M \leq N$ and N are the number of transmit and receive antennas respectively, form a unitary matrix, i.e., one with orthonormal columns. This signal structure was shown to be capacity-achieving for high SNR or $T \gg M$ [HM00]. The capacity expression of the block-fading channel with non-coherent reception has a geometrical interpretation as sphere packing in the Grassmann manifold $\mathcal{G}(T, M)$: the set of all M-dimensional subspaces of $\mathbb{C}^T$ [ZT02]. In particular, the columns of the USTM matrices are viewed as a basis that spans an M-dimensional subspace. At the receiver, the channel modifies the basis but keeps the subspace

unchanged. Therefore, information can be encoded into subspaces and not into the particular basis.

Up to now, PSAM and USTM have been considered as different techniques that require different modulators and demodulators. We can show, however, that there is certain equivalence that can be used to harmonize them.

2.5.2 Constellation structure

The transmitted signal, $\boldsymbol{X}$, in a block-fading channel is selected from a constellation of matrices with $L = 2^{RT}$ elements, where R is the transmission rate. Letting $\mathcal{S}(\boldsymbol{X})$ be the subspace spanned by the columns of $\boldsymbol{X}$, the constellation $\mathcal{C}^{\mathrm{U}} = \{\boldsymbol{X}_i^{\mathrm{U}}\}_{i=1}^{L}$, where $\boldsymbol{X}_i^{\mathrm{U}*}\boldsymbol{X}_i^{\mathrm{U}} = \boldsymbol{I}_M$ is the $M \times M$ identity matrix, and $\mathcal{S}(\boldsymbol{X}_i^{\mathrm{U}}) \neq \mathcal{S}(\boldsymbol{X}_j^{\mathrm{U}})$ for all $i \neq j$, is a USTM. In the case of a PSAM, a certain quantity of channel uses is reserved to carry the pilot signals. If optimization over the training and data powers is allowed, the optimal number of training channel uses is M [BH03]. In this case, $T - M$ channel uses are used for data transmission, in which matrices from a coherent code $\mathcal{A} = \{\boldsymbol{A}_i\}_{i=1}^{L}$, where $\boldsymbol{A}_i \in \mathbb{C}^{T-M\times M}$, are transmitted. A PSAM can be constructed concatenating the training and data signals in $\mathcal{A}$. However, it is also possible to allow these signals to spread throughout the entire coherence interval by means of a unitary matrix. In particular, let $[\boldsymbol{B}_1 \quad \boldsymbol{B}_2]$ be a $T \times T$ unitary matrix, where $\boldsymbol{B}_1$ and $\boldsymbol{B}_2$ are the matrices with the first M columns and the last $T - M$ columns, respectively. The constellation $\mathcal{C}^{\mathrm{P}} = \{\boldsymbol{X}_i^{\mathrm{P}}\}_{i=1}^{L}$, where

$$\boldsymbol{X}_i^{\mathrm{P}} = \boldsymbol{B}_1\boldsymbol{P} + \boldsymbol{B}_2\boldsymbol{A}_i, \tag{17}$$

and $\boldsymbol{P} \in \mathbb{C}^{M\times M}$ is a pilot matrix, is a PSAM. The pilot matrix can be retrieved as $\boldsymbol{B}_1^*\boldsymbol{X}_i^{\mathrm{P}} = \boldsymbol{P}$, and the data as $\boldsymbol{B}_2^*\boldsymbol{X}_i^{\mathrm{P}} = \boldsymbol{A}_i$.

2.5.3 Equivalent constellations

In this section we describe equivalent PSAM and USTM constellations in terms of the spanned subspaces.

2.5.3.1 USTM equivalent to a PSAM

It can be shown that, if $\boldsymbol{P}$ is invertible and $\mathcal{C}^{\mathrm{P}} = \{\boldsymbol{X}_i^{\mathrm{P}}\}_{i=1}^{L}$ is constructed from $\mathcal{A}$ as in (1), then $\boldsymbol{A}_i = \boldsymbol{A}_j$ if and only if $\mathcal{S}(\boldsymbol{X}_i^{\mathrm{P}}) = \mathcal{S}(\boldsymbol{X}_j^{\mathrm{P}})$. In other words, the subspaces spanned by PSAM matrices constructed from different coherent signals, $\boldsymbol{A}_i$ and $\boldsymbol{A}_j$, are different, and hence, a PSAM could be decoded detecting the transmitted subspace, like in a USTM receiver. It is important to note

that no linear independency is assumed between $\boldsymbol{A}_i$ and $\boldsymbol{A}_j$, i.e., even if $\boldsymbol{A}_i$ and $\boldsymbol{A}_j$ are linearly dependent, the addition of the invertible pilot matrix makes $\boldsymbol{X}_i^{\mathrm{P}}$ and $\boldsymbol{X}_j^{\mathrm{P}}$ linearly independent.

Thanks to this, any PSAM has an equivalent USTM in terms of the spanned subspaces. The elements of the USTM can be obtained from the elements of the PSAM by either a QR decomposition, the Gram-Schmidt process, or any other orthonormalization.

2.5.3.2 PSAM equivalent to a USTM

The equivalence also exists in the other direction. In particular, it can be shown that, if $\boldsymbol{P}$ is invertible, for each USTM $\mathcal{C}^{\mathrm{U}} = \{\boldsymbol{X}_i^{\mathrm{U}}\}_{i=1}^{L}$, there is a $T \times M$ unitary matrix $\boldsymbol{B}_1$, with orthogonal complement $\boldsymbol{B}_2$, such that $\boldsymbol{B}_1^* \boldsymbol{X}_i^{\mathrm{U}}, i = 1, \dots, L$, are invertible and $\{\alpha \boldsymbol{X}_i^{\mathrm{U}} \boldsymbol{R}_i\}_{i=1}^{L}$ is a PSAM whose elements span the same subspaces than the elements of $\mathcal{C}^{\mathrm{U}}$, where

$$\boldsymbol{R}_i = \left(\boldsymbol{B}_1^* \boldsymbol{X}_i^{\mathrm{U}}\right)^{-1} \boldsymbol{P}, \tag{18}$$

$$\alpha = \sqrt{\frac{LM}{\sum_{i=1}^{L} \mathrm{tr}\left(\boldsymbol{R}_i^* \boldsymbol{R}_i\right)}}. \tag{19}$$

Using this result, it can be shown that the USTM can be decoded using the equivalent PSAM, whose pilot matrix is $\alpha \boldsymbol{P}$ and coherent code is $\mathcal{A} = \{\boldsymbol{A}_i\}_{i=1}^{L}$, $\boldsymbol{A}_i = \alpha \boldsymbol{B}_2^* \boldsymbol{X}_i^{\mathrm{U}} \boldsymbol{R}_i$.

2.5.4 Harmonized decoders and performance evaluation

As highlighted in the previous section, a USTM can be decoded with the equivalent PSAM, and a PSAM with the equivalent USTM. This allows for a good harmonization of both transmission methods, i.e., coherent PSAM and non-coherent USTM, in which they can be received with any of the two corresponding reception methods. In this sense, a receiver that implements a coherent reception method can be used to receive PSAM- or USTM-based transmissions.

The price to pay is that the optimum receiver – based on Maximum-Likelihood (ML) – has certain performance loss if the equivalent, and not the original, constellation is used in the reception phase. This loss can be, however, negligible in some cases, as depicted in Figure 2-22. The figure shows the constellation matrix detection error for $R = 2$, $T = 4$, $M = N = 2$, and four combinations of transmitters and receivers. In particular, the figure shows the performance of a transmitter using a USTM and a receiver using the same USTM, a transmitter using a USTM and a receiver using the equivalent PSAM, a transmitter using a PSAM and a receiver using the equivalent USTM, and a transmitter using a PSAM and a receiver using the same PSAM. The USTM has been constructed using the method in [GD09], and the coherent code of the PSAM corresponds to spatial multiplexing, i.e., independent symbols in each antenna and

channel use, from a 4-QAM modulation. The equivalent constellations were constructed following the indications in the previous section. The results of Figure 2-22 show an SNR shift of 0.5 dB when the equivalent PSAM, instead of the original USTM, is used by the receiver, and of only 0.05 dB when the equivalent USTM is used.

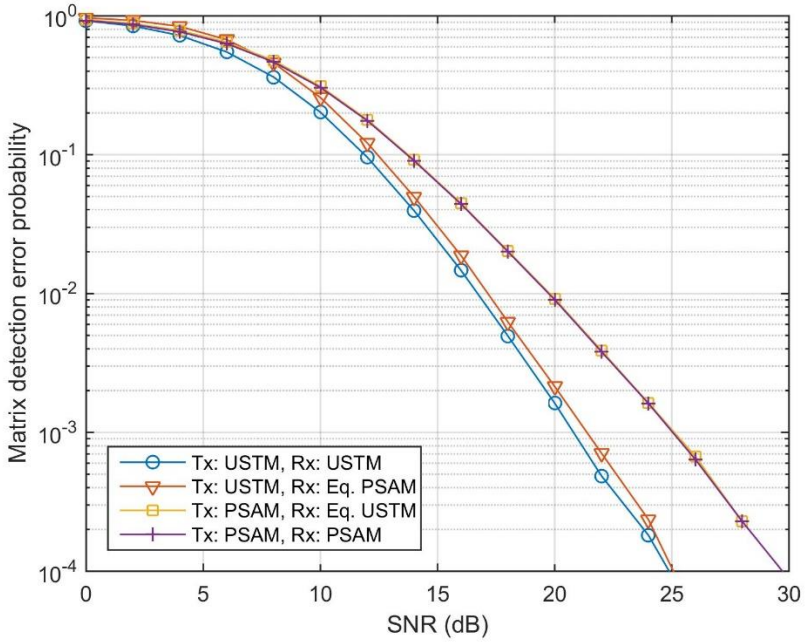


Figure 2-22: Matrix detection error probability of USTM and PSAM constellations detected using the same constellations and their equivalent counterpart.

2.6 Considerations for higher layers

This section centres on the investigation of implications of AIV harmonization on the design of a scalable and forward compatible UP by building on existing AI harmonization considerations in [MET216-D41]. Work captured in [MET216-D41] is a starting point, with further input from agreements captured in [MET216-D22], with LTE-A as a benchmark for simplifying certain functions, combining functions of different layers, as well as introducing new functionalities. The focus will be on resource allocation to explore different options as well as the challenges and requirements imposed to upper layers by multi-numerology and service multiplexing.

Firstly, 5G should efficiently support Dynamic Resource Allocation (DRA) of different numerologies in FDM/TDM fashion. As aforementioned, FDM allows more flexibility in scheduling UEs that are configured with different numerologies. Guard tones can be inserted between frequency resources to reduce the hardware complexity due to the strict requirement of OOBE. The resource allocation can be easily limited to happen only in the resources of each individual numerology. For TDM, since only a single numerology exists per TTI, resource

allocation can be confined within each individual numerology in each TTI in TDM. Furthermore, cross numerology resource allocation is also applicable for a single service when necessary. Whether or not to apply cross-numerology resource allocation will affect the design of upper layer, in particular the MAC design.

Secondly, multiplexing of multiple services imposes another challenge to UP functionality design. For example, xMBB and uMTC services can be multiplexed and dynamic resource allocation and sharing should be supported between xMBB and uMTC. uMTC transmission may occur in resources scheduled for ongoing xMBB traffic. There are several approaches proposed including:

- *Pre-emption-based multiplexing*: since the base station (BS) cannot predict when uMTC data is arrived, rate matching in LTE may not be proper for pre-emption. The other approach is to puncture some resources of already scheduled xMBB by just replacing xMBB data in specific resources with uMTC data. Issues such as indication of uMTC information, timing of the indication delivery need to be considered in order to recover xMBB data transmission.
- *Superposition-based multiplexing*: in certain particular scenario, it might be very difficult and burdensome that BS provides puncturing information to xMBB receiver. Although xMBB and uMTC receivers can suffer from interference effects incurred by superposition transmission, they can overcome the effects by applying interference suppression techniques such as Minimum Mean Square Error Interference Rejection Combining (MMSE-IRC), symbol-level Interference Cancelation (IC), interference aware detection, and other similar techniques. In addition, the BS can reduce the interference effects by controlling superposition parameters such as the number of superposed symbols, power ratio between superposed symbols, and similar means.

The additional information and procedures needed in multiplexing services have implementation implications on upper layers in the sense that they will affect aspects such as signaling design and measurement procedure, etc.

2.7 Conclusions on PHY layer aspects enabling multi-service support

In this chapter, we have looked at main enablers that will allow the 5G AI to support the strongly heterogeneous 5G services. In particular, these are our main results:

- In Section 2.2, we have looked at new waveforms that are being studied for their good performance in some of the new 5G services. We defined a harmonized implementation framework in Section 2.2.3 that allows a harmonized implementation of multiple of these waveforms. This implementation presents some complexity (related to chip space) vs. latency (related to chip clock speed) trade-off that has been analysed in Sections 2.2.3.3 and 2.2.4.

- In Section 2.3, we analysed the suitability of the wide range of bands that will be available for 5G for the three service types (xMBB, mMTC, and uMTC). In particular, we proposed the use of below 6 GHz bands for all services in order to provide good coverage, and above 6 GHz bands to provide the required capacity for xMBB. We also motivated a potential use of above 6 GHz bands for uMTC, and more specifically, for V2V communications.
- In Section 2.4, we looked at the case in which only one band is available and, hence, all services have to be supported in it. Due to the strong differences among services, we concluded that, in such a case, different numerologies and frame structures have to co-exist in the same band. We analysed potential problems and provided guidelines to efficiently use of the frequency resources in this scenario.
- In Section 2.5, we presented a mathematical framework that describes both coherent and non-coherent communication techniques in such a way that they can be considered equivalent. We concluded that both techniques can be completely harmonized, in the sense that a receiver designed for coherent reception can be used to receive signals in a non-coherent manner and vice versa.

These conclusions can be used to give a tentative answer to some of the METIS-II key RAN design questions. More specifically, with respect to the AIVs expected to be introduced or evolved from existing standards, an AIV for below 6 GHz is expected to be an evolution of current 4G standards, at least from a UP design point of view. Only the case of D2D communications may require special waveforms to counteract the effects of asynchronicity. Above 6 GHz new AIVs with special frame structures may be required to, e.g., manage massive MIMO and channel estimation. Vehicular communications, especially for road safety, may require new AIVs to efficiently deal with multicasting, asynchronicity and reliability.

With respect to different AIVs that devices should support, this will depend on the purpose of each device. The harmonized 5G AI should allow that purpose-specific devices implement only necessary functionalities. For instance, a laptop thought to be in a static or quasi-static indoor environment should implement AIVs for below and above 6 GHz with massive MIMO support and high-order modulations. On the other hand, an in-car communication unit for V2V communications should implement an AIV with high reliability and multicasting features. Hence, this in-car unit may not require high-order modulations and massive MIMO support.

3 Service-oriented functional UP design

This chapter discusses design considerations for a service-tailored 5G UP architecture. This involves the functional / protocol design which has to be adapted to meet 5G service diverse requirements, and in particular extreme Mobile Broadband (xMBB), massive Machine Type Communications (mMTC), and ultra-reliable Machine Type Communication (uMTC). The high-level view of the design considerations, which are discussed in this chapter are illustrated in Figure 3-1.

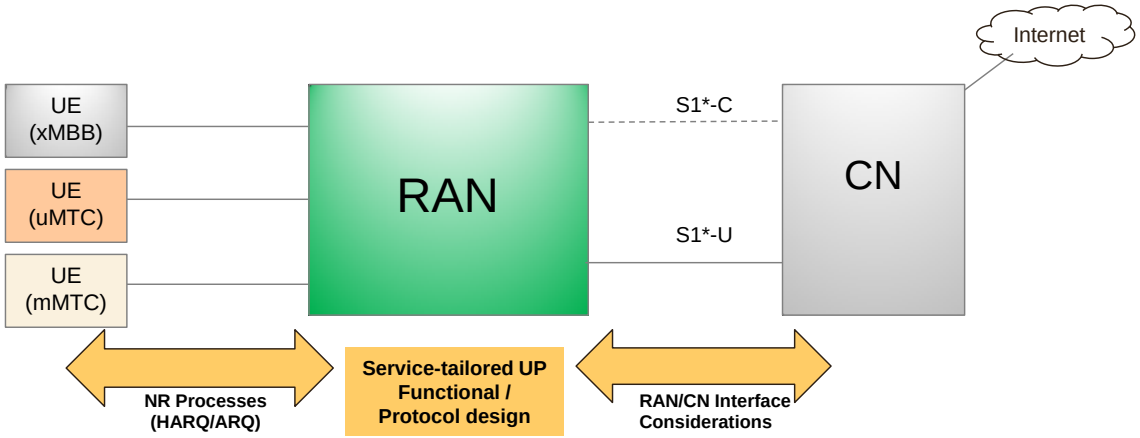


Figure 3-1: High level view on Service-oriented UP design

As a first step towards capturing the current flat UP design, Section 3.1 presents the state-of-the-art air interface UP protocol design and key challenges. Thereafter this chapter proposes some candidate solutions for enhancing the protocol stack in a service-oriented manner. One key consideration in NR is the multi-connectivity support assuming multiple AIV and multi-level designs. In this context, one key challenge is how to perform certain key RAN processes (e.g. HARQ/ARQ) in multi-level multi-AIV scenarios with different functional splits and service requirements. This is a necessary step towards developing the UP design, since this will affect the protocol / functional configurations; hence will be discussed in the chapter as a key enabler for NR design.

To this end, one important aspect is the RAN/CN interface, which will provide requirements to the UP design (e.g. the UP design might need to be adapted in order to be able to meet different interface options, given the physical deployment). On the other hand, the service requirements might necessitate the use of certain RAN/CN Interfacing options to be able to meet certain KPIs (e.g. low end-to-end latency). To this end, Section 3.3 discusses different protocol options for the RAN-CN interface as well as new QoS model destined for efficient QoS

flow to bearer mapping. Finally, Section 3.4 discusses the 5G standardization status and agreements on UP design and Section 3.5 summarizes the key findings of this chapter.

3.1 Air interface UP Protocols and Functions: Overview of main challenges

Based on 3GPP LTE [3GPP15-36300], the UP protocol design and the underlining RAN functions can be summarized at the following figure (Figure 3-2).

PDCP	Header compression and decompression;
	AS Security
	PDCP re-establishment for RLC AM including In-sequence delivery and Duplicate detection
	Timer-based SDU discard in uplink
	Retransmission of PDCP SDUs at handover and split bearer
RLC	RLC TM for system information, paging and SRB0;
	Concatenation, segmentation and reassembly
	RLC UM includes: Reordering, Duplication detection, RLC SDU discard
	RLC AM including all RLC UM functions also includes: Error Correction through ARQ Re-segmentation
	RLC re-establishment.
MAC	Mapping between logical channels and transport channels; Multiplexing/de-multiplexing
	Error correction through HARQ
	Scheduling (Scheduling information reporting, Priority handling between logical channels of one UE; Priority handling between UEs by means of dynamic scheduling)
	Transport format selection;
	Random access
	DRX in connected mode
PHY	Error detection, correction
	Transmission techniques (multi-carrier, modulation, MIMO)
	Synchronization (time, frequency)
	Measurements
	RF processing

Figure 3-2 Conventional protocol / functional design in LTE

The key challenges with the air interface protocol / functional design – focusing on the entire protocol stack as shown in Figure 3-2 above – are:

PHY/MAC:

- In 5G, PHY and Medium Access Control (MAC) configurations will need to be specific to service characteristics in many scenarios. If multiple services are active in a UE, this UE

cannot schedule logical channels only using prioritized bit rate as it is currently done in LTE. For example, if the UL grant is for xMBB then uMTC packet should not be scheduled in this grant as uMTC packet transmission requirement cannot be met by the MAC layer configuration (e.g. HARQ) and physical layer configuration (e.g. numerology) for xMBB. So, some mechanism (other than prioritized bit rate as used in LTE) may be needed to map the UL grant to one or more logical channels.

- As explained in Chapter 5 in [MET216-D41], the modelling of a MAC entity (as single or multiple MAC entities) is an open issue. In solutions where a single MAC scheduler cannot by design handle packets from multiple numerologies, multiple MAC entities may be required. However, a single MAC entity seems entirely appropriate if we assume a single MAC scheduler is used to handle packets from multiple numerologies. This solution has the benefit of forward compatibility, meaning that there is no need to add a new MAC entity whenever a new numerology is introduced.
- One other key outstanding issue is how to enable MAC to differentiate between different numerologies (the modelling of MAC entity will rely heavily on this). 3GPP has recently agreed that for multiple numerologies in PHY, at least the TTI length of the numerology(s) will be visible to MAC; however which (if any) additional characteristics of the numerology are also visible to MAC is an open research area.
- The current version of LTE HARQ protocol can be fast; however not reliable enough to meet 5G requirements. In order to compensate that, Radio Link Control (RLC) is required to ensure reliability which renders additional latency. To be able to meet 5G requirements, the HARQ protocol has to be much faster, with lower overhead, be more reliable, and operate on a flexible timing base.
- In 5G (unlike in LTE), HARQ parameters may need to be configured differently for different services. Similarly, not every access scheme is applicable for all services. Therefore, radio-bearer specific Layer 2 (L2) configuration in 5G may need to include the configuration of HARQ parameters and multiple access schemes.

RLC/PDCP:

- Current Packet Data Convergence Protocol (PDCP) / RLC layering in LTE might not be flexible enough to meet diverse KPIs. Functionalities like redundant header processing, re-ordering and duplication detection in both layers might provide additional overhead and might not be required by all services.
- Moreover, in case of multi-connectivity (which is a very key scenario in 5G e.g. with below and above 6GHz AIVs); different functional / bearer split options will introduce new challenges regarding the interaction of functions. For example, if retransmissions are handled in RLC and PDCP and PDCP has multiple RLC entities, there should be coordination for re-transmissions in different protocol entities.

3.2 Air Interface protocol configuration and functional mapping for multi-service support

This section discusses different studies related to Service-Oriented UP Design. In Section 3.2.1, the requirement for new functionalities due to NR characteristics and the service requirements is discussed in detail. Moreover, an enhanced air interface protocol stack solution, which aims to

decrease processing delay in a service tailored fashion, is described in Section 3.2.2. Finally, key candidate solutions for re-transmission handling in NR are presented for multi-connectivity scenarios and for harmonization among services (Section 3.2.3).

3.2.1 Requirement for novel NR Functions

Multi-AIV Support: In NR, each AIV can be characterized by different sets of physical layer features (waveform, multiple access scheme, frame structure, numerology, etc.). Assuming multiple AIVs as part of NR, the AIV-to-resource mapping is a key control functionality closely coupled with UP design. More elaboration on the control functions related to AIV-to-Resource Mapping can be found in [MET217-D62].

- **Semi-static AIV to Resource mapping:** In this case, the mapping between AIV and spectrum is fixed for a given time period and is performed in PDCP. RAN allocates AIVs to time/freq. resources according to Radio Resource Control (RRC) or Operations, Administration & Maintenance (OAM) settings. However, the mapping can be updated based on the load, traffic volume, and radio condition changes. In this case, backhaul is not required to be ideal between BSs supporting different AIVs.
- **Dynamic AIV to Resource mapping:** In each TTI**,** time/freq. resources can be allocated dynamically per AIV in MAC. This is ideal for high dynamic scenarios such that radio resource requirements for each AIV are very dynamic accordingly. MAC layer assigns radio resources to each AIV, and further determines the time/frequency resource allocation to services using certain AIV at per TTI basis. For this case, service multiplexing level can be high with proper Radio Resource Management (RRM) algorithms; nevertheless this requires ideal backhaul between nodes (or co-location of nodes) supporting different AIVs.

Support for high frequencies (mmWave): mmWave radio is envisioned as a key enabler for providing high capacity in 5G RAN. However, due to the path loss limitations in high frequencies, directional antennas with adaptive beam-forming and overlapping coverage are going to be used to exploit the benefits of operation in mmWave radio and ensure high coverage. In the mmWave radio, the main challenges regarding UP design are the following:

- High penetration loss of mmWave frequencies can severely deteriorate the performance and hence, maintaining reliable connectivity is a challenge especially for delay critical services.
- Wireless channel conditions and link quality can change significantly during movement of users, calling for fast RRM decisions and multi-connectivity support. User mobility also causes significant and rapid load changes and handovers due to small coverage areas of access nodes. Therefore, connection management and load balancing in conventional RRM functionalities need to be revisited to cope with the aforementioned challenges [LPV+17].
- Due to highly directional transmissions, crosslink interference characteristics become much different from sub-6 GHz systems. For example, there can be flashlight effects (an interfering beam hits a user).

From UP design point of view, it is recommended to adapt MAC design for high directivity. In 3GPP, beam management [3GPP16-RP163476] is discussed as a key functionality to meet the capacity and coverage requirements by performing resource-to-beam allocation and beam sweeping / steering at MAC / Lower Layer. The aim of the beam management is to maintain connectivity between UE and a serving access node during mobility and radio environment change. From a set of candidate beams, it would be dynamically decided in which beams each user can connect to and what should be the beam-to-resource allocation. To this end, one important aspect is the RAN/CN interface options which will provide requirements to the RAN design (e.g. the RAN design might need to be adapted in order to be able to meet different interface options, given the physical deployment). On the other hand, the service requirements might necessitate the use of certain RAN/CN Interfacing options to be able to meet certain KPIs (e.g. low end-to-end latency).

3.2.2 Enhancements to existing protocol stack

Due to the stringent latency requirements, as imposed by certain 5G services, the air interface protocol processing delay is a key consideration before re-visiting the protocol design, especially for multi-level Ultra Dense Networks (UDNs) with various multi-connectivity options. In Table 3-1, the contributions of different L2 functions to the overall processing delay are exemplified for 3GPP LTE system. We can observe that PDCP introduces major contribution to L2 latency, but also functions like header processing overall represent considerable contributions.

In LTE, traffic can be handled differently per bearer in RLC/PDCP layers, but some functions are more static with limited flexibility, e.g. different RLC modes, Robust Header Compression (RoHC) profiles. In MAC/PHY traffic is multiplexed across bearers given the channel prioritization and other constraints. Hence, air-interface protocols can be potentially grouped in “per bearer” and “across bearer” protocols. Given that in 5G services, RLC functionalities can be either moved together with PDCP (e.g. for MTC) or together with MAC (due to ARQ timing constraints), an alternative layering could be envisioned as a candidate solution to allow for service-tailored optimization of functionalities.

Table 3-1: L2 Latency Contributions in LTE [SSH+09]

Sub-layer	Function	Overall Latency contribution
PDCP	ROHC	20.01%
	De-ciphering	59.16%
	Header processing	7.83%
RLC	Reassembly	8.60%

	Re-ordering	0.40%
	Header processing	1%
MAC	De-mux	0.84%
	Header processing	2.16%

The proposal of a two-layer approach introduces a) the Lower Layer with traditional MAC and flexibly some RLC functions given the service requirement and b) the Upper Layer which includes traditional part of RLC, PDCP and some modified functions. In this two-layer proposal, processing gains can be achieved, in terms of latency, due to less header processing, single duplication detection and re-ordering, modular re-transmission control (instead of MAC, RLC, PDCP, two levels of re-transmission are proposed). This architecture can better deal with multi-connectivity and internal RAN functional split (Upper layer can be centrally deployed and have multiple independent Lower layer entities (corresponding to different AIVs).

Below, we present some key considerations for the configuration of functions, which can be tailored for the three main 5G service types.

- In xMBB, in order to accomplish the ultra-high throughput KPI, a key consideration is to enable the operation in higher frequencies, which can offer much higher bandwidth. Furthermore, high centralization of UP functions, as well as coordinated multipoint transmission, are envisioned as key technologies to meet the target KPIs.
- On the other hand, uMTC is a service type which requires ultra-high reliability and low latency for time critical use cases (e.g. vehicular safety, eHealth). For achieving ultra-high reliability, multi-connectivity / carrier aggregation (CA) can be seen as key technology enabler. In particular with the proposed layering, the upper layer will be able to connect to multiple lower layer entities; thus potentially lowering the processing delays (taking also into account the re-transmission processes). To this end, due to the small and fixed size of the packets, function like segmentation is not required; and fixed size Logical Channel Prioritization (LCP) can be used at Lower Layer.
- In mMTC service type, one of the main KPIs is to maximize the connection density. Since, the traffic requirement is low, the devices are expected to be static (e.g. sensors) and the density is expected to be high, group-based functionalities (see also [MET217-D62]) both in lower and upper layer are envisioned. Further, some functions related to mobility can be disabled.

Figure 3-3 illustrates some exemplary functions in the enhanced two-layer protocol stack. Some functions can be seen as control logics closely coupled with the legacy UP functions. More analysis of the presented control functions can be found in [MET217-D52] [MET217-D62].

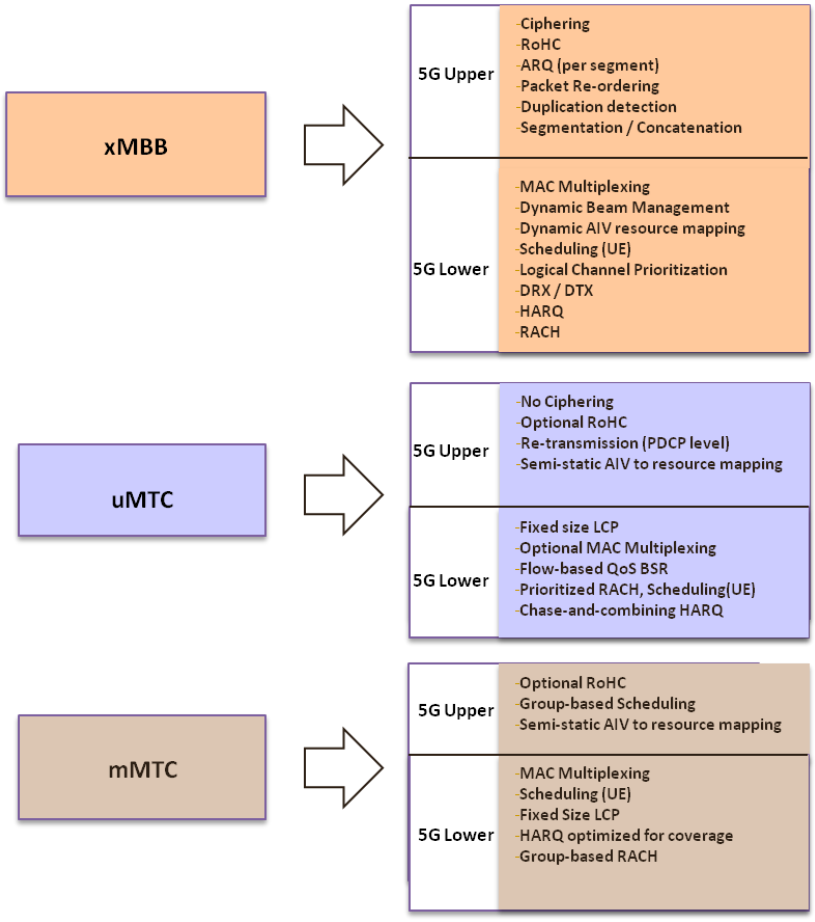


Figure 3-3: Exemplary Service tailored U-Plane Protocols/Functions in NR

In Figure 3-4, we show some indicative comparison on the UP L2 processing latency. The benchmark is the LTE eNB processing latency which is 1ms for the UP [3GPP16-25912]. It is worth mentioning, that the latency strongly depends on factors like the TTI size and the processing capabilities at RAN; hence only a simplified scenario is exemplified to better capture the trend when we enable/disable functions in a service-tailored manner. Moreover, since the functions strongly depend on the RAN/UE characteristics, for the uMTC service type we present two cases: uMTC1 or no/low mobility use cases and uMTC2 for high mobility scenarios (e.g. for vehicular safety, assuming high mobile cars). The configuration for different services can be as follows:

- For xMBB, marginal gain is achieved by having less header processing due to the two-layer approach.
- In mMTC and uMTC1, most Upper Layer (PDCP and RLC) functions are disabled; however we assume an aggregation layer (on top of PDCP) for group-based Upper Layer functionalities. uMTC1 shows higher latency, since it involves also packet re-ordering / assembly functions in Upper Layer.

- For uMTC2, due to the expected frequent handovers and multi-connectivity support, upper layer includes ciphering; however RoHC is not included as well as RLC functions, to further decrease the latency. Also, both uMTC cases assume no multiplexing at Lower Layer / MAC.

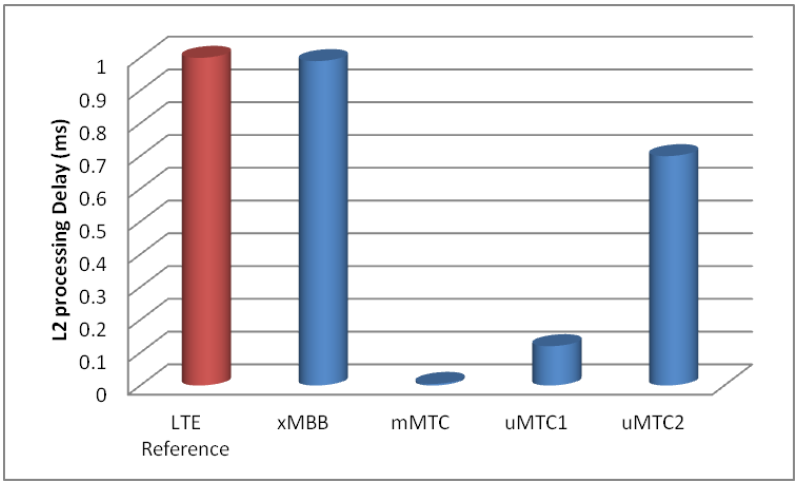


Figure 3-4: L2 UP latency comparison

3.2.3 Enhancements to NR processes

Focusing on MAC functionalities, which require further attention as identified in the introductory chapter and [MET-II16-D41], HARQ is a known PHY/MAC layer scheme used for retransmission. Typically, it includes chase combining or incremental redundancy with a Boolean feedback, i.e.: Acknowledgement (ACK) in case of successful decoding, negative acknowledgement (NACK) in case of failed decoding or no received packet (running out of timer). The HARQ process has round trip time (RTT), which impacts the latency of used service. HARQ RTT depends of various factors such as TTI, ACK/NACK transmission times, UE, eNB processing times etc.

In downlink, eNB can use the 8 HARQ processes in any order asynchronously. UE will not have any information regarding used HARQ process before receiving DL data and is required to be informed about it via Physical Downlink Control Channel (PDCCH). In uplink, this process is synchronous in nature. Here retransmissions are scheduled at fixed time intervals. UE will have to use specific HARQ process at specific sub-frame and eNB will know exactly which HARQ process comes when. UE will have to use the same HARQ process number every 8 subframes. Also, running on the fixed timing causes some other implementation issues. For instance, in unlicensed spectrum scenarios, “listen-before-talk” requirement sometimes does not allow UEs to send HARQ feedback.

These are rigid in 4G (meaning that the timing relationship between the initial transmission and the re-transmission is fixed), hence motivating the flexible design of HARQ to satiate stringent service requirements in 5G. It is therefore of great importance to study asynchronous HARQ operation for both DL and UL (while maintaining optional alternatives, such as e.g. optionally synchronous HARQ operation configured by RRC).

To this end, this section discusses enhanced re-transmission processes for NR. In Section 3.2.3.1, new HARQ/ARQ design is proposed to further exploit the benefits of NR features, like multi-connectivity and multi-hop transmission. Following, in Section 3.2.3.2 a harmonized HARQ framework is discussed designed for multi-service support.

3.2.3.1 Re-design HARQ/ARQ based on modified RLC

Supporting multi-hop and mobility are key features foreseen for 5G systems. Different L2 protocol architectures will impact the multi-hop design, e.g., multi-hop ARQ and routing.

NR HARQ/ARQ for both NR downlink and uplink is proposed in this section based on the structure of RLC ARQ and MAC HARQ in LTE. HARQ is modified to make it faster, with lower overhead, be more reliable and not requiring fixed timing.

3.2.3.1.1 Downlink HARQ/ARQ design

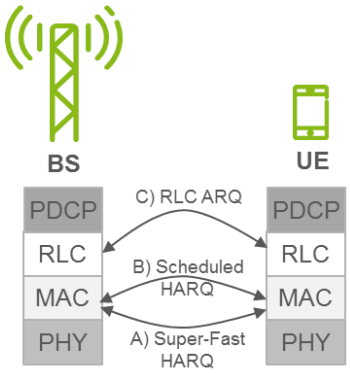


Figure 3-5: Improved ARQ for single hop NR. The HARQ protocol utilizes two different feedback mechanisms: one "Super-Fast" (A) and one "Scheduled" (B). On top of this the RLC layer (C) handles residual errors (e.g. due to mobility) and re-segmentation.

For NR we propose an improved HARQ protocol with two components as illustrated in Figure 3-5: the HARQ protocol utilizes two different feedback mechanisms/processes: one "Super-Fast" (A) and one "Scheduled" (B). On top of this, the RLC layer (C) handles residual errors (e.g. due to mobility) and re-segmentation. In particular:

A. The "Super-Fast HARQ" feedback provides as fast-as-possible HARQ feedback, albeit not fully reliable.

B. The “Scheduled HARQ” feedback provides an efficient, almost-100% robust, and it may be suitable for use in e.g. Dynamic TDD scenarios due to flexible timing.
C. RLC ARQ can be seen as an addition, which is similar to the current LTE RLC Acknowledged Mode (AM) ARQ [3GPP15-36322].

A. Super-Fast HARQ Feedback

The component A is ultra-lean and transmitted as soon as possible. It provides feedback for one or a few downlink transmissions, as shown in Figure 3-6.

The feedback can be indicated by a single bit whether it is ACK/NACK based on a success or a failure of the received downlink assignment decoding. Upon receiving the “Super-Fast HARQ” feedback, either the same data will be retransmitted by the base station or new data on a different HARQ process will be transmitted.

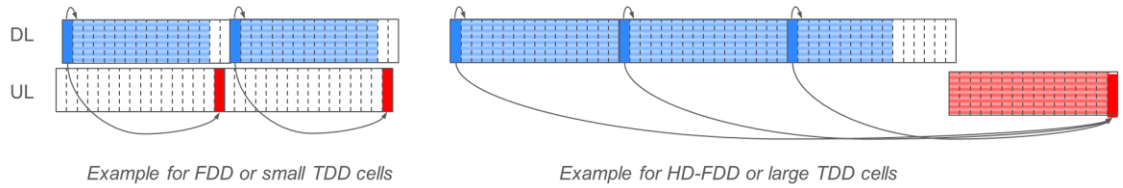


Figure 3-6: Example where the fast HARQ feedback is transmitted at the end of the first available UL transmission occasion. Left: Frequency Division Duplex (FDD) or small cell TDD example where HARQ feedback is included in a single OFDM symbol; Right: example with half-duplex FDD or large cell TDD where the fast HARQ feedback is included in the last OFDM signal of the scheduled uplink transmission.

B. Scheduled HARQ Feedback

The “Scheduled HARQ” feedback (B) is a multi-bit HARQ feedback scheduled on the uplink data channel, typically the direct transmittable Physical Downlink Channel (dPDCH), defined in [MET-II16- D41]).

The second component B is a multi-bit HARQ feedback which is suitable for dynamic TDD where Cyclic Redundancy Check (CRC) is needed for the robustness. The base station can indicate to the UE which HARQ process(es) should be reported in the feedback by sending an Uplink Control Information (UCI). The size of the feedback can be large to include for all the allocated HARQ processes for a User Equipment (UE) in the downlink, and the size can be smaller to contain only a subset of the allocated HARQ processes.

In LTE, 1 bit is used to indicate whether it is ACK/NACK per HARQ process, in NR this can be more than 1 bit per HARQ process. For instance, 2 bits can be used for each HARQ process:

- A New Data Indicator (NDI) toggle-bit, that indicates if the feedback relates to an odd or even packet in the process. This bit toggles each time the UE receives a new-data-indicator in the downlink grant associated with this HARQ process;
- An ACK/NACK-bit for the HARQ process. Moreover, the number of HARQ processes is fixed in LTE. In contrast, new 5G radio interface design can consider the maximum number of HARQ processes and the feedback delay to be configurable.

3.2.3.1.2 Uplink HARQ/ARQ design

In LTE, the HARQ operates in a synchronous mode, meaning that the retransmission occurs in a fixed time interval, i.e., 8 ms after the initial transmission for FDD. The feedback is explicitly communicated by a dedicated control channel, i.e., Physical Hybrid ARQ Indicator Channel (PHICH). This fixed timing is not suitable for 5G, e.g., in an unlicensed operation, all the sub-frames may be not always available due to Listen Before Talk (LBT) limitation.

Therefore, we propose that for 5G uplink HARQ, the HARQ feedback is not explicitly communicated but dynamically handled by allocating uplink grants with the same process ID and a NDI to request retransmissions. This would reduce the overhead of using an explicit control channel for feedback and the network can dynamically schedule the uplink grant, such that the fixed timing issue between the transmission and the corresponding feedback is avoided. A potential drawback of the proposed uplink HARQ is that there can be a false detection of uplink grant by a UE, leading to the UE discarding undelivered data. However, the probability of multiple consecutive false detection events, while having data in the uplink buffer is very small due to the following reasons:

- UE would keep requiring the uplink grant from the network if the buffer is not empty, and
- the network will have much more reliable transmission scheme for the uplink grant if an error is identified, e.g., with a reasonable CRC and search space.

The uplink grant based feedback is more suitable for scheduled uplink transmissions compared to contention based uplink transmissions. In scheduled uplink transmissions, an uplink transmission is always scheduled by the network via an uplink grant. However, for a contention based uplink transmission, the transmission resources are pre-configured and uplink grants do not correspond to every contention based transmission. When designing the HARQ scheme for contention based transmission, one consideration is that the soft combining of retransmission attempt should not be supported. The reason is that the contention based channels can easily collide, since the resource management is not dynamically controlled by a scheduler as well as in LTE UL. Therefore, the soft-buffers are likely to be very noisy and soft combining may not be beneficial.

When transmitting on a contention-based resource, the UE needs to include an additional sequence number which is encoded as an uplink control information element in the uplink data channel. Thus, ARQ without soft-combining should be supported and the ARQ feedback can in that case be provided in a separate feedback message in a MAC control element. However,

typically an uplink contention based transmission will be followed by an uplink grant for a scheduled uplink transmission which then implicitly also contains the ARQ feedback for the contention based transmission.

3.2.3.1.3 Dynamic soft HARQ buffers

In LTE, each HARQ process has a corresponding soft HARQ buffer in the UE, however, the size of the soft buffer is designed to be a UE capability. A UE supporting a certain maximum number of HARQ processes should not be required to also support the same size of soft buffer, especially at very high data rates. As shown in Figure 3-7, soft buffers for large sized packets (e.g., tens of Gbps) can be very large and very expensive. But for medium size packets, the soft buffers are small and the UE should support soft combining in this case.

A UE vendor may decide to put a very large soft buffer in the device, considering the cost-benefit trade-off. The benefit of improving performance via soft packet combining in low rate cell-edge scenarios is usually significant with reasonable cost.

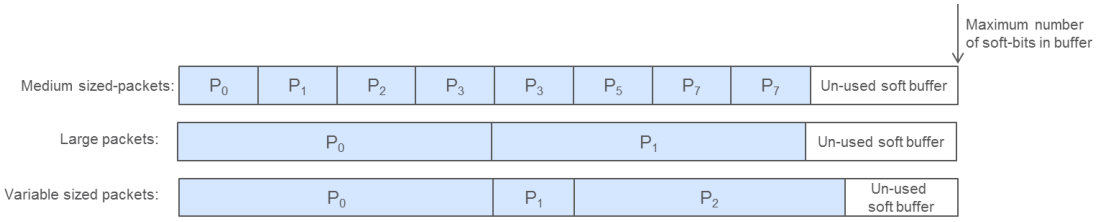


Figure 3-7: The number of HARQ process for which the UE perform soft packet combining may depend on the packet size.

3.2.3.1.4 Multi-hop ARQ protocol architectures

In a multi-hop/self-backhauled scenario, e.g. relay or mesh network, some additional considerations are required in the design of HARQ/ARQ protocol. At first, it is evident that the different hops in a multi-hop/self-backhaul chain may have very distinct characteristics. They may differ in terms of e.g.:

- Radio Link Conditions/Quality (e.g. SINR, channel properties etc.)
- Reception (Rx) /Transmission (Tx) Capabilities (e.g. number of antennas, max Tx power, beamforming, receiver algorithms, interference suppression capabilities etc.)
- Traffic and Routing (e.g. number of multiplexed users, number of multiplexed routes, amount buffering etc.)
- (Dynamic) TDD Configuration etc.

Hence, it is quite reasonable to assume that per-hop RRM mechanisms (i.e. link adaptation, segmentation etc.) are required. Secondly, as the number of hops grows the cumulative probability of failure in the per-hop ARQ mechanism somewhere along the multi-hop/self-backhauled chain increases. Also further, in case of classical mobility (e.g. the UE attaches to

another AP/ Relay Node (RN) - possibly also belonging to another anchor BS) or when the path to the UE is re-routed (e.g. RNs in the multi-hop/self-backhauled chain is removed/added) this needs to be accounted for. Essentially, in a scenario with mobility and/or not fully reliable per-hop (H)ARQ, we need a separate mechanism to ensure end-to-end reliability. In other words, yet another end-to-end ARQ layer is required in these situations, as discussed below. Due to the aforementioned reasons, there are three possible ARQ protocol architectures for the multi-hop/self-backhauled scenarios:

- Alt. 1 “Per hop HARQ/RLC ARQ”: The single-hop ARQ architecture. This indeed violates the assumption on having an end-to-end mechanism discussed above, but is included for further motivation.
- Alt. 2 “End to End RLC ARQ”: Again, the same single-hop ARQ architecture is utilized over each hop as in Alt. 1 above – but now with only HARQ and no RLC over each hop. A higher layer RLC (inclusive of ARQ, segmentation etc.) is instead placed only at the end-point nodes, i.e., in the BS and the UE.
- Alt. 3 “Two Layered RLC ARQ”: This is essentially a combination of the two other ARQ architectures with a single-hop ARQ including HARQ and RLC ARQ for each hop and – in addition – an extra higher layer RLC is placed on top of this in the end-point nodes

The three protocol architectures are illustrated in Figure 3-8. In the end-to-end RLC layer (Alt. 2/Alt. 3), the transmitting RLC entity at one endpoint (BS or UE) would place at the buffer each transmitted packet until it is positively acknowledged by the receiving RLC entity (UE or BS). Thereafter, this packet is removed from the buffer. The transmitting RLC entity needs to set ARQ retransmission timer depending on the total end-to-end delay, in order not to avoid premature retransmissions. An appropriate timer value can be estimated in various ways, but this procedure will obviously be cumbersome in dynamically changing environment and/or complex routing scenarios. In such cases it is better if the timer is disabled and the endpoint retransmissions are triggered only by explicit negative acknowledgements from the receiving endpoint RLC entity (Note: disabling the transmitting RLC entity timer should be only considered when the reliability of negative acknowledgement reception from the receiving RLC endpoint is ensured)

We note that Alt. 1 may not be a suitable candidate in mobility scenarios and might not be fully reliable per-hop (H)ARQ mechanism. Due to that if a packet that fails to be delivered in any hop, then it would be handled by a multi-hop relay ARQ protocol architecture available only in Alt.2 and 3, as illustrated by Figure 3-8.

Furthermore, relying on end-point retransmissions in Alt. 2 is inefficient due to the lack of support of the per hop ARQ. In particular, a failed packet cannot be re-segmented if the radio condition is changed, thus it may require MAC level segmentation (to support per-hop re-segmentation). Hence, the two layered ARQ (Alt. 3) is a more generic architecture to suit the foreseen scenarios.

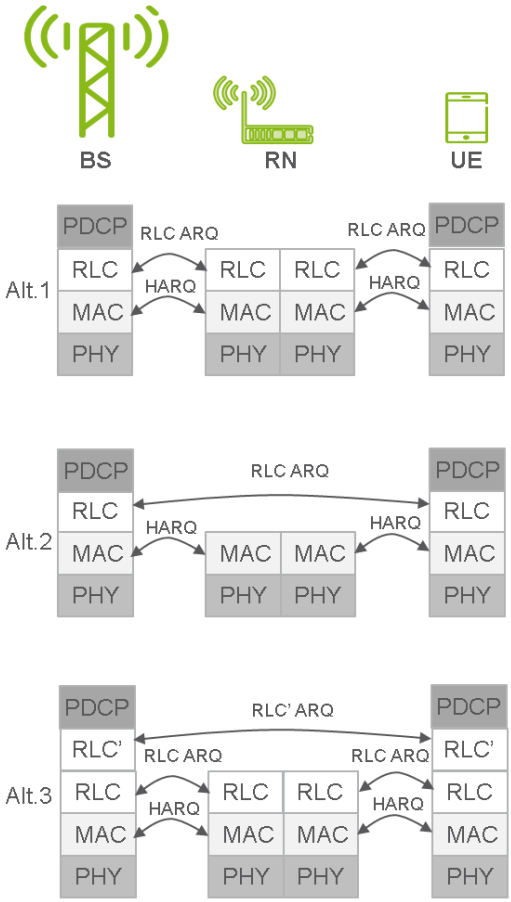


Figure 3-8: The three possible multi-hop/self-backhauled ARQ architectures (including HARQ and RLC ARQ for each hop) and an additional higher layer RLC at the end nodes.

Relay ARQ

Relay ARQ is a modified version of the two layered ARQ architecture (Alt. 3). It integrates the ARQ of the higher RLC layer into the per-hop relay RLC layer, as shown in Figure 3-9.

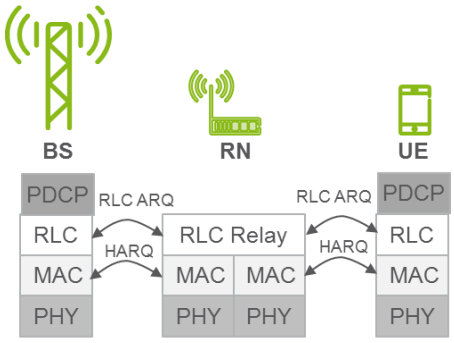


Figure 3-9: Multi-hop Relay ARQ protocol architecture.

The key idea of relay ARQ is that the temporary retransmission responsibility is delegated from the sender node (the source node or the relay node) step-wise from node to node until finally the data is received at the receiver, i.e., the retransmission happens only in the hop causing the failure. The ultimate retransmission responsibility, however, remains at the source node (BS or UE) to maintain end-to-end reliability. No matter whether the two Layered ARQ approach of Alt. 3 or if the relay ARQ architecture is used, it is only in the endpoints (e.g. BS and UE) where in-order delivery of RLC SDUs shall be employed. The reason for this is that it is only the higher protocol layers in the end-points, which may require in-order delivery of data, whereas requiring in-order delivery in the intermediate nodes may risk under-utilizing the links. Also, by not requiring in-order-delivery in each intermediate node, we also allow for the possibility to map the data packets over multiple-paths, and hence achieve a load balancing over intermediate links and nodes.

3.2.3.1.5 Performance Evaluations

We evaluate the three multi-hops ARQ protocols (Alt. 1, Alt. 2, Alt. 3), including the relay ARQ approach denoting it as Alt. 4. Mainly, we investigate the delay incurred by employing the four protocols in different situations, for example, different number of hops and different radio link qualities. A poor radio link quality can result in erroneous feedback, like the NACK-to-ACK error, or the transmission error (i.e., retransmission occurs). In the simulation setup, we model these two types of errors with NACK-to-ACK error probability and initial transmission error probability. For the traffic model, we model the packets arrival rate as fixed rate and the size of the packet is kept fixed for simplicity. For the protocol layer, we mainly model the RLC layer, which delivers the incoming packet with a variable delay. The variable delay consists of a fixed delay and an additional delay per required HARQ transmission attempt. The fixed delay for each packet is 4 ms, and if there is a HARQ retransmission, additional delay is added which is in our model equal to 8 ms. We model the link layer as a dummy HARQ model for simplicity, in which two probabilities are used for modelling the link quality. One is the error rate probability for each HARQ process transmission, the other is the probability for NACK-to-ACK feedback error. Typically, in our simulation, the error rate probabilities for each HARQ process transmission are {0.3, 0.1, 0.003, 0.001}. We vary the NACK-to-ACK error probability between 0 and 0.1 for different simulation cases discussed below.

Note that, in our evaluations, there is not any noticeable difference in the delay performance of per-hop RLC (Alt. 1) and relay RLC protocols (Alt. 4). This is due to the fact that the scenarios with mobility and relay failures are not evaluated. In general, we expect relay RLC protocol to be superior than the per-hop RLC.

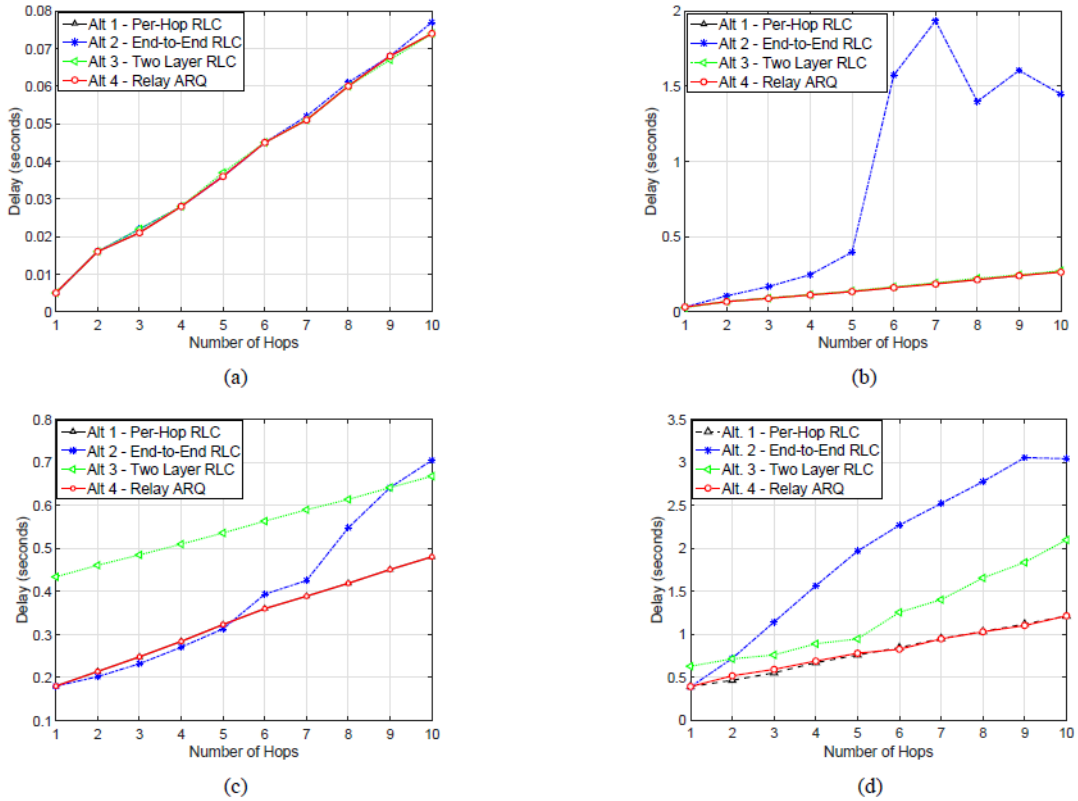


Figure 3-10: (a) The delay performance for different hops, with initial transmission error probability but no NACK to ACK error; (b) The delay performance for different hops, with initial transmission error probability as well as NACK to ACK error; (c) The delay performance for different hops, link quality is the same with (a) but the traffic is in high load; (d) The delay performance for different hops, link quality is the same with (b) but the traffic is in high load.

In Figure 3-10 a) and d), the delay performance of the considered protocols is shown as a function of number of hops under different error probabilities and traffic loads. Each point in these figures corresponds to the 50% point in the Cumulative Distribution Function (CDF) for the specified number of hops. In a), the initial transmission error probability is 0.3 and there is no NACK-to-ACK error. We observe that there is no significant difference in the performance of different protocols; the increasing delay is mainly due to the increasing number of hops. In Figure 3-10 b), NACK-to-ACK error probability is set to 0.1. We observe that the end to end RLC ARQ protocol is very inadequate, especially when the number of hops is increased. This happens due to the fact that a NACK-to-ACK error will trigger the retransmission from the end node no matter in which hop the error actually occurs. The end-point retransmissions are expensive and should be avoided. In Figure 3-10 c) and Figure 3-10 d), we repeated the simulations with the error probabilities assumed in Figure 3-10 a) and Figure 3-10 b) respectively, under a high traffic load. High load traffic refers to the situation that there are always available packets in the buffer (full buffer traffic scenario). According to c), the two layered protocol (Alt. 3) experiences large delay, because two distinct layers add much more overhead than the single layer protocols. Again, we notice that as the number of hops

increases, end-point retransmissions become more costly. In Figure 3-10 d), we can observe again that the end-to-end RLC ARQ protocol is very poor due to the fact that NACK-to-ACK errors trigger the end node retransmissions which are very costly. Worthwhile to mention that in Figure 3-10 the black curve (Alt. 1) and green curves (Alt. 3) are overlapping with the red curves (Alt. 4).

3.2.3.2 Harmonized HARQ for multi-service support

A flexible and enhanced HARQ design which is configurable as per service is proposed. The HARQ timings are proportional to scalable TTI sizes. The number of stop and wait (SAW) channels are allowed to be configurable per user per link direction with asynchronous HARQ operation in both UL/DL. Further, asymmetric UL/DL operation is preferred with respect to used TTIs.

- HARQ timings proportional to scalable TTI
  - Dynamically scalable TTI sizes for each scheduling instant of a user
    - For latency sensitive data, scheduling in short TTIs (uMTC)
    - For larger data loads, scheduling in longer TTIs (xMBB)
- Asymmetric UL/DL operation
  - UL & DL coverage is asymmetric.
  - UL requires longer transmission time for coverage challenged terminals.
  - DL could still serve such users with short TTIs (due to higher transmit power)
  - Transmission times for sending UL and DL information must be allowed to be set differently.

3.2.3.2.1 Downlink HARQ RTT

Figure 3-11, shows the HARQ operation in DL. The BS transmits the data in DL, which is received and decoded by UE. On successful reception and decoding, an ACK is sent, otherwise a NACK is transmitted in UL. The HARQ RTT is measured from the time of transmitting the scheduling grant and payload in DL, until the BS can begin a new transmission or corresponding retransmission over the same SAW channel.

The DL HARQ RTT is measured as $T_{rtt-DL} = T_{DL-tx} + 2T_P + T_{UE} + T_{A/N} + T_{eNB}$, where T_{DL-tx} is DL transmission time, T_P is propagation time, T_{UE} is UE processing time, T_{eNB} is BS processing time and $T_{A/N}$ is ACK/NACK transmission time in UL. Time alignment is assumed at BS and UE, considering T_P.

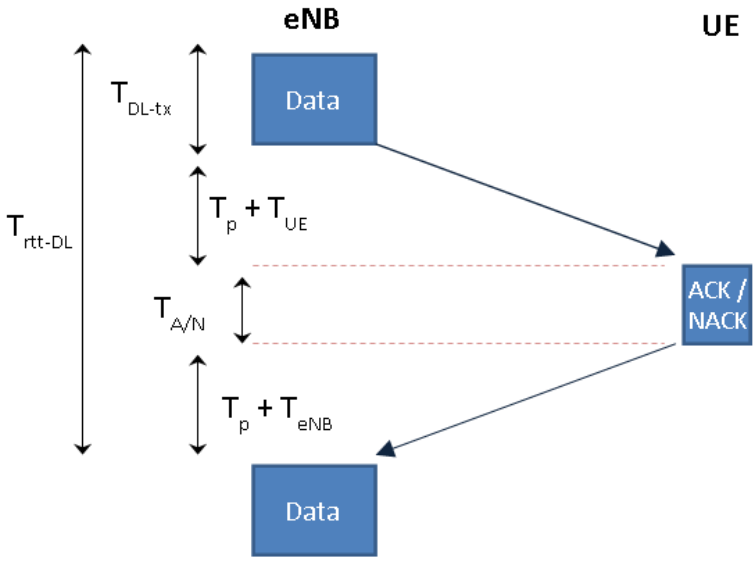


Figure 3-11: DL HARQ RTT

3.2.3.2.2 Uplink HARQ RTT

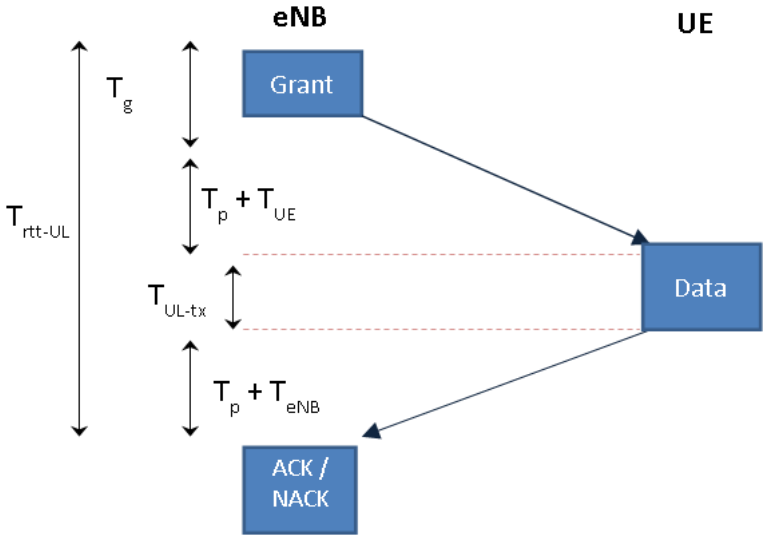


Figure 3-12: UL HARQ RTT

Figure 3-12 demonstrates the UL HARQ operation. The scheduling grant is transmitted by BS in DL. Upon this, data is transmitted by UE in UL. ACK/NACK is sent by BS correspondingly. The UL HARQ RTT is measured from the time the grant is transmitted from BS until start of subsequent ACK/NACK transmission at BS.

The UL HARQ RTT is measured as $T_{rtt-UL} = T_{UL-tx} + 2T_P + T_{UE} + T_g + T_{eNB}$, where T_{UL-tx} is UL transmission time, T_P is propagation time, T_{UE} is UE processing time, T_{eNB} is BS processing time and T_g is grant duration in DL. It is assumed that transmit and receive TTI are time aligned at BS and timing advance is applied to the UE TTI alignment, i.e., UE transmit and receive TTIs start T_P before and after BS TTI respectively. Thus effect of T_P is nullified.

3.2.3.2.3 Number of SAW channels

The number of SAW channels must be sufficiently large to keep constant flow of data transmission. However larger number of SAW channels leads to larger HARQ buffer to store received packets in case of failure. Thus, there is a trade-off between the SAW channels and the buffer size. The minimum number of SAW channels required to avoid HARQ stalling is given as $N_{SAW} \geq T_{rtt}/T_{DL-tx}$, to allow continuous transmission in DL, number of SAW channels needs to be selected based on TTI size in DL, along with bundling feedback in UL [KBP+16]. For instance, with $T_{DL-tx} = 0.2$ ms and $T_{UL-tx} = 1$ ms, UE must bundle 5 feedback messages and transmit them at once. i.e ACK, if all 5 HARQ processes are received correctly, NACK otherwise. If bundling is not used then there will be less active HARQ processes affecting throughput.

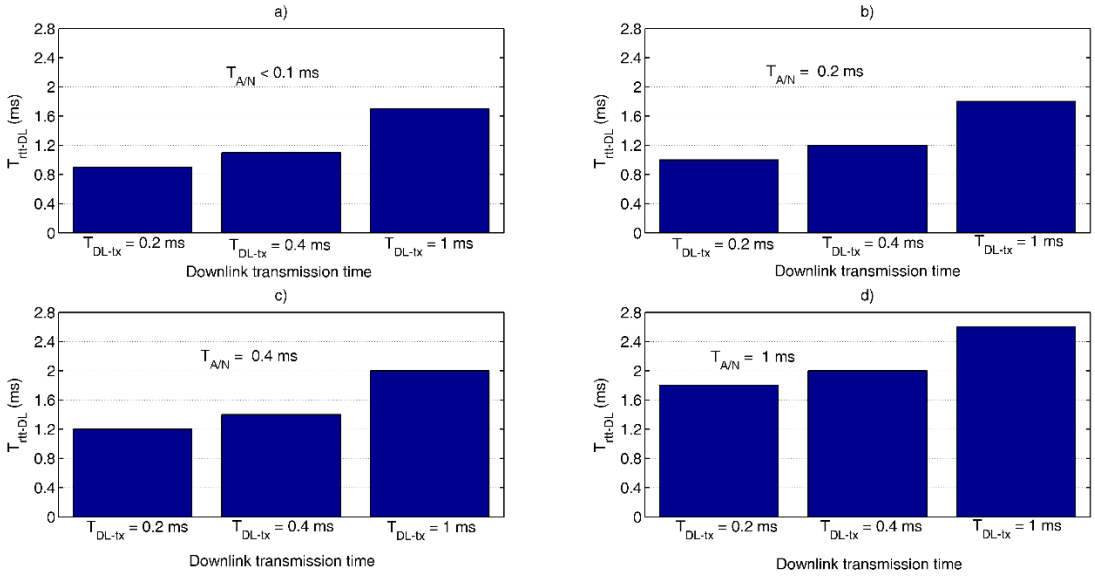


Figure 3-13: Downlink HARQ RTT

The key concept is to configure the parameters such as TTI size, number of active SAW channels and enable asymmetric UL/DL HARQ timings and operation corresponding to service requirements (e.g., latency). Further, the extent of harmonisation of HARQ for all three service types is analysed and framework for harmonised HARQ is provided. The variations of DL and

UL HARQ RTT with different values of UL/DL transmission times, ACK/NACK transmission times are analytically evaluated and results are discussed below.

Figure 3-13 shows DL HARQ RTT for various values of T_{DL-tx} and $T_{A/N}$. TTI is assumed to be 0.2 ms and processing times at UE and BS (T_{UE}, T_{eNB}) are assumed to be 0.3 ms [KBP+16]. Figure 3 a) shows the case, where acknowledgement time in UL is less than 0.1 ms. In such case, RTT target of 1 ms for uMTC can be met (given $T_{DL-tx} < 0.4$ ms). In case of $T_{A/N}$ and TTI being 0.2 ms (figure 3 b)), RTT increases to 1.2 ms (given $T_{DL-tx} < 0.4$ ms). Even in such cases, RTT can be reduced to 1 ms by exploiting schemes for early feedback calculation [BKP+16]. It should be noted that minimum DL HARQ RTT can become as long as 2.8 ms, considering TTI of 1 ms and $T_{A/N} = 1\,ms$. This implies a low coverage UE with longer TTI transmission in DL. The minimum DL HARQ RTT in LTE is 8 ms, the reduction observed here in comparison is due to flexibility in both frame structure and HARQ timing.

Figure 3-14 depicts the UL HARQ RTT for various values of grant duration (T_g). Even before the DL TTI is fully received, UE can start processing resource grant right after receiving it. This leads to reduced RTT in UL. The RTT target of 1 ms for uMTC cases can be met in UL for $T_g < 0.1$ ms and $T_{UL-tx} = 0.2$ ms (indicating a UE in good coverage). For a UE in low coverage (far away from eNB) with $T_{UL-tx} = 1$ ms, UL HARQ RTT increases to at least 1.8 ms.

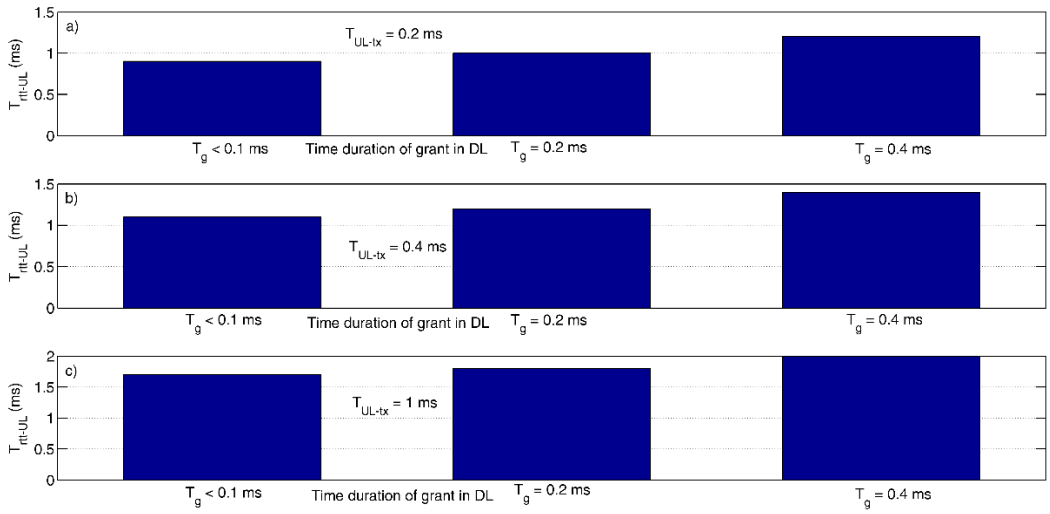


Figure 3-14: Uplink HARQ RTT

Table 3-2 shows the comparison between numbers of SAW channels for different asymmetric UL/DL timings. The UL transmission time can limit the number of SAW channels, since UE can transmit only one ACK/NACK at a time in UL. The column in grey shows the number of SAW channels when bundling is used. The entry in red shows the stalling of HARQ process when single A/N is used instead of bundling, leading to lesser number of active HARQ processes.

Table 3-2 Configurable number of SAW channels

T_{DL-tx} ms	UL transmission time: $T_{A/N}$ ms							
	< 0.1		0.2		0.4		1	
0.2	5	5	6	6	4	7	2	10
0.4	3	3	4	4	4	4	3	6
1	2	2	2	2	3	3	3	3

Table 3-3 shows the key functionalities in HARQ design for each service type and their target KPIs. These functionalities need to be configured as per link/service requirements. xMBB requires higher throughput/spectral efficiency, sometimes even at cost of latency. Thus, scalable TTI is preferred, which can be even longer depending on constraints. Avoiding stalling of HARQ processes is another key aspect for xMBB services. This can be facilitated by enabling HARQ feedback bundling. Further, instead of having single bit ACK/NACK multi-bit richer feedback will allow improving throughput further.

Table 3-3 Key functionalities for harmonised HARQ

Harmonised HARQ framework		
xMBB	**uMTC**	**mMTC**
• Scalable TTI • ACK/NACK bundling • Larger HARQ buffer • Multi-bit feedback [SKD13] *Higher throughput, spectral efficiency*	• Scalable TTI (shorter) • Multi-bit feedback • Early feedback [BKP+16] *Lower latency*	• Scalable TTI (longer) • Smaller HARQ buffer • Small Packet HARQ process *Larger coverage, energy efficiency*

In case of uMTC, latency is the key focus. Scalable TTI and HARQ timings are essential, preferably shorter. Richer feedback will help in reducing latency by allowing more information about how close the receiver was at decoding the failed HARQ transmission [SKD13]. Early feedback will allow the prediction of success of decoding in advance to send out ACK/NACK. For mMTC, main concern is energy efficiency and coverage. The TTI can be longer compared to uMTC and xMBB. Small packet HARQ process is to be supported with keen focus on energy efficiency.

3.3 RAN/CN Design Considerations

3.3.1 5G QoS model considerations

In the current LTE model [3GPP15-36300] all packets with same QoS requirements use one radio bearer (defined on Uu radio interface between UE and eNB – see Figure 3-15), without differentiating between the traffic flows, which may cause unfairness and/or cause unnecessary delays e.g. for services when packet importance varies (e.g. due to a different coding rate). In such situation, it may be beneficial to differentiate QoS level based on the packet significance. Current LTE QoS model does not support such differentiation. Also, 5G technology will face new use cases where different AIVs may need to be prioritised or different services will be run in parallel, for which different packet treatment may need to be enforced. Previous radio access technologies such as Universal Mobile Telecommunications System (UMTS) and LTE relied heavily upon an idea of the radio bearer, which can be viewed as a logical concept embracing several principles and mechanisms that allow for classifying packets and providing the corresponding treatment (e.g. scheduling, queuing, rate shaping, RLC settings). In LTE radio bearer is mapped to EPS bearer which could be either Guaranteed Bit Rate or Non-Guaranteed Bit Rate and include parameters such as QoS Class Identifier (consists of priority, acceptable delay and packet loss rate) and Allocation and Retention Priority.

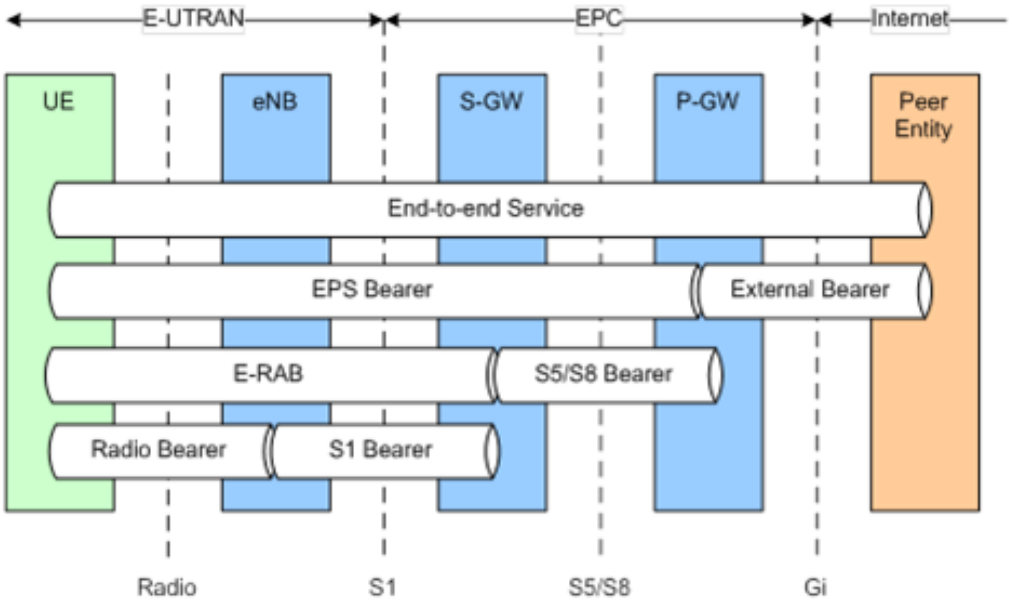


Figure 3-15: Bearer service architecture in LTE

Due to new AIVs and simultaneous support for diverged service requirements (not supported in LTE), new QoS framework for 5G needs to be more flexible (considering both RAN and CN aspects) to allow better handling of different traffic flows (e.g. packets with different importance) with the same legacy QoS requirements. One issue is that nowadays, packets with the same or

no specific requirements are placed into the radio bearer / queue, which in turn might impact negatively performance of other flows in the same queue. One of the most typical examples is when a TCP connection request (or response) packet is placed to the end of queue thus delaying connection setup time. Another typical problem is message bursts coming from different flows, which are put to the end of queue thus delaying each other instead of being fairly scheduled in parallel.

To address these issues, 5G needs to consider a model in which the core network can detect and classify packets with the same QoS requirements (e.g. bit rate, delay, priority, reliability) belonging to different flows with finer granularity (i.e. below Evolved Packet System (EPS) bearer which is the QoS granularity level in the legacy LTE architecture). As a result, introduction of new lower level "QoS flow" definition could be considered. Furthermore the tradeoffs in different mapping approaches between QoS flows and RAN radio bearers need to be analysed e.g.:

- One-to-one – establishing a separate RAN radio bearer for each QoS flow may be impractical for efficiency reasons
- Many-to-one – i.e. different EPS flows with same QoS requirements can be multiplexed to one RAN radio bearer (there could be flows with strict QoS requirements where the possibility to map particular flows to a separate radio bearer may be beneficial).

Alternatively, simultaneous availability of both mapping options could be considered, but there is a trade-off between increased user plane flexibility and complexity in this approach – e.g. number of supported radio bearers (with associated queues and user plane state machines) in UE may be limited due to its hardware capabilities and supported service types.

All these new mapping approaches may require new QoS flow definition considering different 5G services and also further analysis which user plane stack layer would optimally implement such flow / bearer mapping and what additional QoS requirements and signalling would be required e.g.:

- QoS flow and its precise definition as a new QoS granularity level
- Association between QoS for DL packets (CN assigned) and corresponding UL packets (UE assigned) and how it is indicated (in-band vs. control plane)
- EPS DL/UL QoS identifiers handling in RAN
- Default QoS rules in RAN

As an example, Figure 3-16 presents four QoS flow to RAN radio bearer mapping solutions. The green colour refers to the traffic flow with same QoS requirements (QoS flow), blue colour refers to the radio bearer, and the red colour is the core network tunnel (e.g. GPRS Tunnelling Protocol (GTP) tunnel on S1 interface in LTE).

For instance, option 1 "LTE" is the legacy LTE solution with one to one mapping of the EPS bearer (virtual connection between the UE and Packet Data Network Gateway (P-GW) with

specific QoS attributes) and RAN radio bearer with no further classification into specific QoS flows (in LTE model, eNB creates the binding between a radio bearer and core network tunnel (S1 bearer) in both the uplink and downlink).

Option 2 “New QoS flow only in RAN” introduces finer granularity QoS flow in RAN.

Option 3 “New QoS flow in RAN/CN” addresses the problem of a separate radio bearer per flow by effectively aggregating flows with the same QoS requirements into one radio bearer. Furthermore, option 3 assumes that the CN follows the legacy QoS model with additional per-flow marking; from the RAN perspective, the gNB will take incoming packets, account for the per-flow marking assigned by the core network, put packets with the same requirements into the one radio bearer.

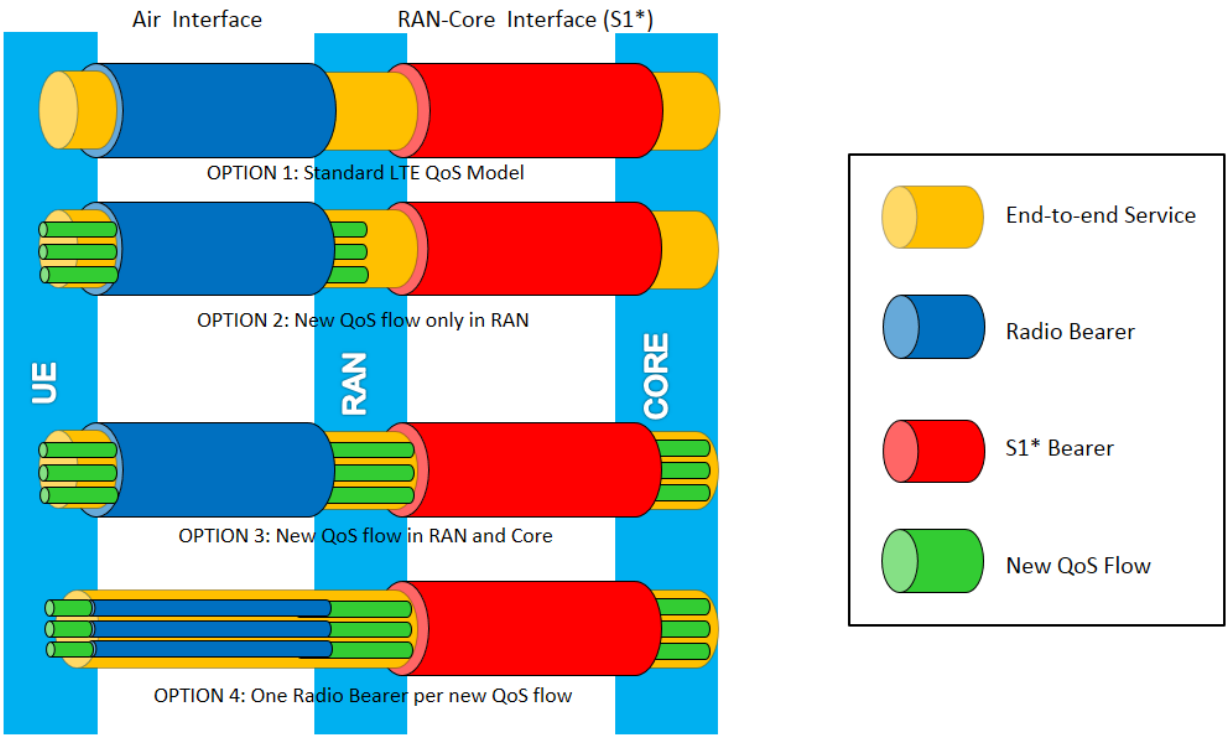


Figure 3-16: Options to map the new QoS flow to the 5G system

Option 4 “New QoS flow only in CN” in Figure 3-16 presents an approach where each QoS flow is mapped to a separate radio bearer (one to one mapping), regardless of flow tunnelling in the core network. The advantage of this solution is the possibility to reuse existing and standardized RAN components, such as PDPC and RLC entities, by effectively allocating a separate queue for each QoS flow in order to protect it from other instances. At the same time, having a separate PDCP/RLC state machine for each flow might create an overhead as it may be difficult

to predict how long a particular PDCP/RLC state machine will be needed; not mentioning the fact the establishment of each radio bearer requires additional control signalling.

In summary, the 5G QoS model and corresponding user plane functionality needs to be further studied and designed to support new AIVs and broad scope of new service requirements allowing efficient and more flexible handling of various types of user data (e.g. by introducing finer QoS granularity). These aspects should be jointly considered in RAN and CN taking into account new 5G framework (e.g. network slicing). Some 5G QoS aspects are also considered in [MET217-D52].

3.3.2 Thin pipes vs. fat pipes on the RAN-Core interface

To provide end-to-end connectivity in today's 4G networks, user and service specific tunnels are setup between the nodes of the EPS. As illustrated in Figure 3-17, user plane tunnels are set-up for each user, for each service type and for each traffic direction (downlink and uplink). Because these tunnels only carry the traffic of a single bearer, they are denoted as "thin pipes". For the set-up, modification or release of these tunnels, a control plane protocol like GTP-C or S1-AP is used.

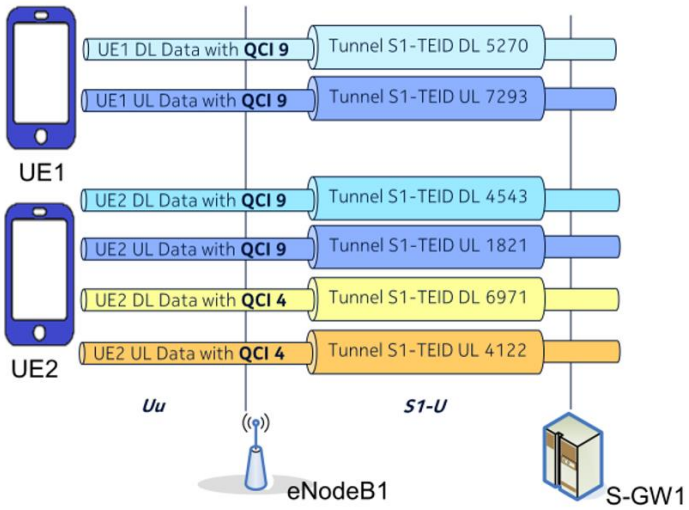


Figure 3-17: Thin pipe tunnelling: One tunnel per user per service type and traffic direction

In contrast to the user and service specific "thin pipes" which carry only packets of a single user, so-called "fat pipes" carry packets of multiple users belonging to the same service (e.g. same Packet Data Network/Access Point Name (APN)) and having compatible QoS characteristics [VHM12]. The principle of aggregating multiple bearers in these fat pipes is illustrated in Figure 3-18. In today's networks, such fat pipes are e.g. used on the 3GPP SWi interface where a WLAN AP aggregates the traffic from all connected devices towards the Trusted Wi-Fi Access

Gateway (TWAG). For 5G, fat pipes can be used on the S1*-U interface as well as on other network internal interface carrying user plane like the X2* interface or core network internal interfaces.

The fat pipe model improves scalability by reducing the number of tunnels. Further, using these fat pipes as static tunnels, the amount of control signalling can be drastically reduced as there is no need for a new tunnel establishment each time a UE uses a new bearer [GZ16], [MET-II17-D62].

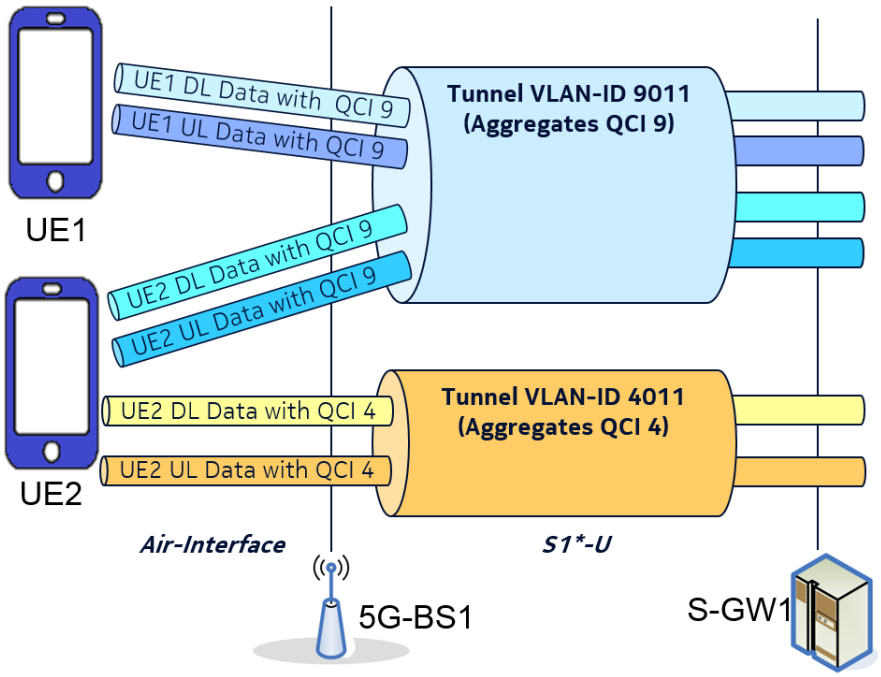
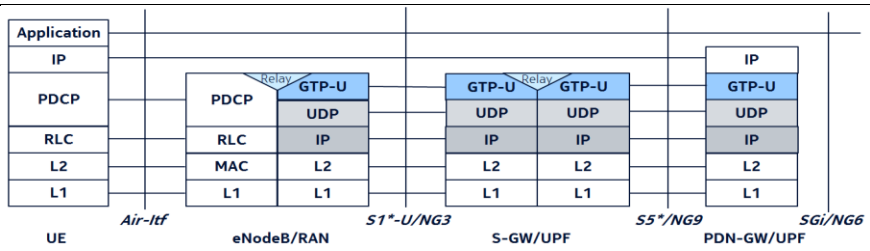
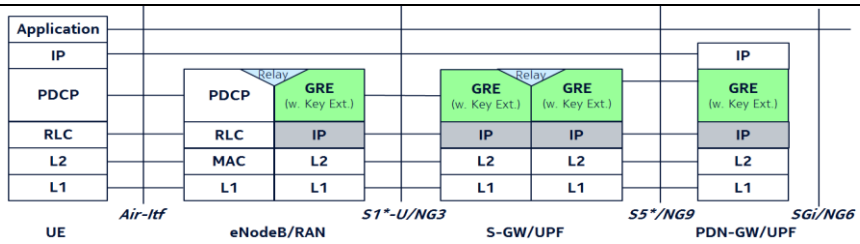
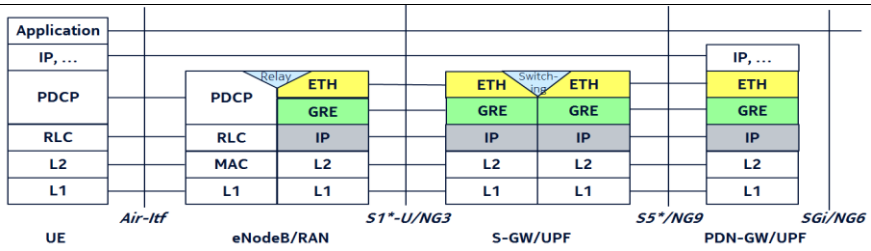
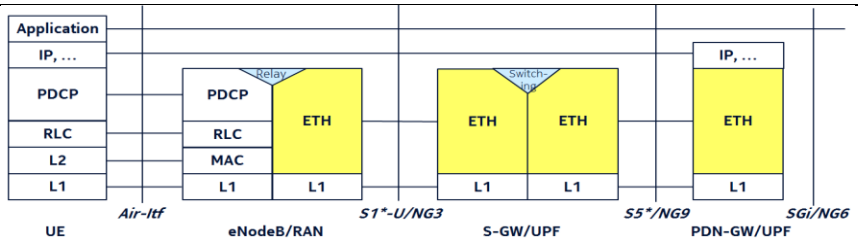
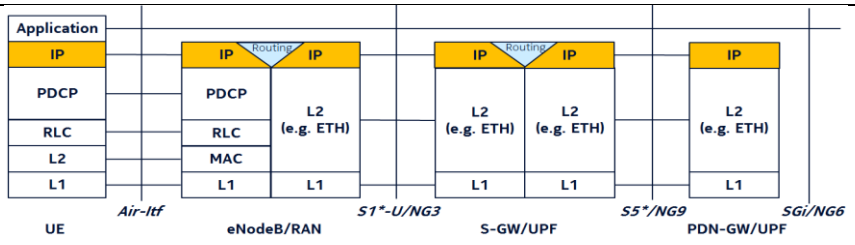


Figure 3-18: Fat pipe tunnelling [GZ16]: Aggregating traffic from different users but with similar service characteristics in one tunnel

3.3.3 Protocol options for the RAN-CN interface

For 5G, it is expected that different user plane protocols can be used on the S1*U-interface, also known as NG-3 interface, between the 3GPP RAN and the Core. This section presents five protocol options for the user plane interface between the 3GPP RAN and Core. The five options (GPRS Tunnelling Protocol for User Plane (GTP-U), Generic Routing Encapsulation (GRE), Ethernet over GRE (EoGRE), L2 datagram (e.g. Ethernet) switching, L3 packet (e.g. Internet Protocol - IP) forwarding) are shown in Table 3-4. Please note that these protocol options can also be applied to other interfaces than the S1*-interface, especially to the user plane interfaces inside the core network (S5*/NG9 interface).

Table 3-4: User plane protocol stack options for the network internal interfaces

Option A: GTP-u/UDP based tunnelling	
Option B: GRE based tunnelling; GRE with header extensions	
Option C: EoGRE: Layer 2 (Ethernet) datagram switching over GRE	
Option D: Layer 2 (Ethernet) datagram switching	
Option E: Layer 3 (IP) packet forwarding	

Below some further explanation on the aforementioned options is presented:

- **Option A** “GTP-U” uses the same protocol stack as in the current S1-interface. This option is based on the thin pipe tunnelling principle. GTP-U is described in [3GPP16-29281] and is transported over User Datagram Protocol (UDP)/IP. Tunnel Endpoint Identifiers (TEID) are used to identify a tunnel. The TEID is assigned by the receiving

endpoint of a tunnel. Thus, via a control plane protocol like GTP-C, the sender must first request a TEID from the receiving endpoint.

- **Option B** "GRE" uses the GRE protocol as specified in [IETF-RFC2784]. The tunnel endpoint IDs can be transported in the key field. Therefore, the "Key and Sequence Numbering Extension" as described in [IETF-RFC2890] is required. Both Option A and B use one tunnel per user per service type and traffic direction and are thus using the thin pipe tunnelling principle as illustrated in Figure 3-17.
- **Option C** "EoGRE" aggregates the traffic of multiple UEs but with similar service characteristics in one GRE tunnel. This option thus uses the fat pipe principle as illustrated in Figure 3-18. The UEs are identified by an IEEE MAC address which is assigned by network. One possibility is to map a 3GPP identifier like the System Architecture Evolution (SAE) Temporary Mobile Subscriber ID (S-TMSI) into the IEEE MAC address as described in [GZ16]. In case the UE used multiple 3GPP based network interfaces in parallel, a Network Interface Identifier (NIID) must also be mapped into the IEEE MAC address. QoS class identifiers (QCIs) can be mapped into the Virtual Local Area Network Identifier (VLAN-ID) used inside the Ethernet header.
- **Option D** "Ethernet datagram switching" simplifies the previous option by using the Ethernet layer without any additional tunnels. As in the previous option, locally administered IEEE MAC addresses are used to identify the 3GPP network interface of a UE. For scalability reasons of the backhaul transport network, methods and protocols like Transport Interconnection of Lots of Links (TRILL) or Shortest Path Bridging (SPB) should be used.
- **Option E** "IP packet forwarding" uses IP forwarding techniques. Typically, the forwarding tables are updated with Software Defined Network (SDN) methods, e.g. when a new UE attaches or in case of mobility.

The following table compares the 5 user plane protocol stack options presented previously:

Table 3-5: Comparison of user plane protocol stack options for the network internal interfaces

	Option A GTP-U/UDP tunnelling	**Option B GRE based tunnelling**	**Option C Ethernet over GRE**	**Option D L2 (Ethernet) datagram switching**	**Option E L3 (IP) packet forwarding**
Tunnelling /Forwarding type	Thin pipe tunnelling	Thin pipe tunnelling	Fat pipe tunnelling	L2 Forwarding	L3 Forwarding
Supported end-to-end protocol	Currently only IPv4 and IPv6. Others only if signalled via NG2-AP	Any	Any	Any	IPv4, IPv6

Header size in bytes (without end-to-end protocol header)	40 (GTP/UDP/IPv4) 60 (GTP/UDP/IPv6)	36 (GTP/UDP/IPv4) 56 (GTP/UDP/IPv6)	42 (GTP/UDP/IPv4) 62 (GTP/UDP/IPv6)	18 (Eth)	18 (L2, e.g. Eth)
Flow (or Flow group) identification	32-bit TEID	32-bit Key	MAC@ + Prio+VLAN-ID	MAC@ + Prio+VLAN-ID	IP@ + TCP/UDP Port
Access type	3GPP only	Any	Any	Any	Any
Already used in 3GPP	Yes: S1, S5, S8, X2	Yes. S5, S8, S2a	No	No	No
Applicability in SDN	Possible	Yes	Yes	Yes	Yes
Mobility	GTP-based	PMIP	SDN-based	SDN-based	SDN-based
Comment	Current standard on S1-interface. Also assumed as default for 3GPP R15.	Similar performance and efficiency as Option A	Used already in Trusted WLAN Access Network (TWAN). Support of Fat-Piping. Better Scalability than Option D (ETH)	Pure Ethernet. UE Addressing with IEEE MAC addresses e.g. based on S-TMSI. Similar to Option E (EoGRE)	Pure IP forwarding. Independent of underlying layer. Compatibility with 3GPP QoS Models FFS. High routing complexity.

Below, the key take-away points of this analysis are highlighted:

- **Option A** “GTP-U” is well suited for standard 3GPP use cases as this protocol stack is already used in 4G systems on the S1-interface as well as on the X2, S5 and S8 interface.
- **Option B** “GRE” has many similarities to option A. GRE is used in 3GPP for non-3GPP access (S2a-interface) and as alternative to GTP on S5/S8 interface inside the Evolved Packed Core (EPC). As GRE is standardized in Internet Engineering Task Force (IETF), this option is also used in non-3GPP networks. Options A and B have similar

performance and bandwidth requirements. As the GRE header has a protocol type field, it easily supports different end-to-end protocols.

- **Option C** "EoGRE" supports fat pipe tunnelling and thus reduces the required amount of control plane signalling. Therefore, it is well suited for mMTC use cases or mMTC slices. It also supports any end-to-end protocol.
- **Option D** "Ethernet" forwards the user data packets on Layer 2 and does not use any tunnelling technique. It has a low protocol header overhead, however, it has a high routing complexity in the network. Further on, it supports IP- as well as non-IP data delivery.
- **Option E** "IP" forwards the user data packets on Layer 3 and does not use any tunnelling technique. It has a low protocol header overhead, however, it has a high routing complexity in the network. It does not support non-IP data delivery.

The usage of tunnelling in options A to C results in a low routing complexity in the transport network: The transport network is responsible for forwarding IP packet data from the CN to the RAN (for downlink and vice versa for uplink). In the case of tunnelling the target address of the IP packets is an address of the base station, which limits the number of hosts in the transport network to the number of base stations (typically thousands). A drawback of tunnelling is the overhead due to the large header size.

Options D and E rely on pure Ethernet / IP in the transport network. This reduces overhead and supports integration with fixed networks (fixed mobile convergence) or future wireless networks (so called forward-compatibility). However, the absence of tunnelling causes UEs to appear as nodes in the transport network, increasing the number of nodes by orders of magnitudes. Growing routing complexity in the elements of the transport network is the result.

It is expected that the first solution supported in 5G is option A "GTP-U" due to its backwards compatibility. Option B "GRE" can also easily be supported in addition or as alternative as it provides similar functionality. For the efficient support of certain services like massive Machine Type Communication, at least one alternative solution should be supported. In order to support IP- as well as non-IP data delivery, Option D "Ethernet" would be well suited. In case of scalability issues with Option D, Ethernet can also be transported via GRE fat pipe tunnels as described in Option C "EoGRE".

3.4 5G Standardization Status on UP Design

In March 2016, 3GPP has started a new study item 'Study on NR Access Technology' [3GPP17-38804] in the Radio Access Network Working Group. The objective is to develop NR access technology to meet a broad range of use cases including enhanced mobile broadband, mMTC, uMTC, and additional requirements including support for frequency ranges up to 100 GHz. The resulting normative specification would occur in two phases: Phase I (planned completion in June 2018) and Phase II (planned completion in December 2019). As part of this

work, the following NR user plane related agreements were captured in Annex A2 of [3GPP17-38804] in Feb 2017:

- **PDCP** - Complete PDCP PDUs can be delivered out-of-order from RLC to PDCP. RLC delivers PDCP PDUs to PDCP after the PDU is reassembled. PDCP reordering is always enabled if in sequence delivery to layers above PDCP is needed.
- **RLC** - NR UP protocol stack supports maintaining of multiple parallel "logical channels" that can be configured with different characteristics and priorities. The ARQ will be supported in RLC.
- **RLC** adds an RLC SN. RLC AM supports T-reordering like functionality for the purposes of determining the content of the RLC status report. The gNB should have means to control which logical channels the UE may map to which numerology and/or TTIs with variable duration. Segment offset based segmentation can be considered for both segmentation and resegmentation as a baseline in NR user plane to support high data rate.
- **MAC** - It is preferable for NR to support only asynchronous HARQ in UL and DL. MAC sub headers placement with respect to the MAC payload can be determined once the rest of the U-plane is more stable. The eNB should have means to control which logical channels the UE may map to which numerology and/or TTIs with variable duration. A UE can support multiple numerologies from a single cell.

Furthermore Table 3-6 summarizes latest agreements (Feb 2017) related to the NR UP design in 3GPP (based on [3GPP17-38804] and [3GPP17-TSG_WG2]).

Table 3-6: Summary of latest 5G (NR) user plane design decisions in 3GPP (Feb 2017)

User plane function	5G New Radio 3GPP standard agreement
PDCP	• Separate Sequence Number in PDCP and RLC (rather than common) • For URLLC (uMTC) services, packet duplication is supported for both user plane and control plane in PDCP for reliability
RLC	• RLC PDUs concatenation for RLC PDUs performed in MAC • ARQ performed on any numerologies / TTI lengths that logical channel is mapped to
MAC	• MAC sub-headers interleaved with MAC SDUs • MAC control elements not placed in the middle of MAC PDU but at beginning or end • UE shall not send padding if data available and remaining Transport Block size greater than predefined number (X) of bytes (in LTE, X=7 bytes) • Single logical channel mapped to one or more numerology

	/ TTI duration • Single MAC entity supports one or more numerology / TTI durations
New QoS framework	• New Packet Data Association Protocol (PDAP) agreed on top of PDCP supporting all QoS functions ○ Supports QOS flow – data radio bearer routing, QoS Flow ID marking in DL / UL packets ○ Single entity for PDU session for all cases connecting to 5G core network
Dual Connectivity (DC) between LTE and NR	• Split bearer support via Secondary Cell Group only for LTE-NR dual connectivity when LTE is master mode
RAN-Core interface	• For the first phase of 5G, only GTP-U based tunneling is supported.

3.5 Conclusions on Service-oriented functional user plane design

This chapter provided an overview on UP design considerations in NR, tailored to different services with tight and diverse KPIs (e.g. latency, reliability, throughput, connection density). Initially, the overview of LTE protocol architecture was presented and the key challenges were highlighted. Furthermore, some key candidate considerations for UP design were discussed regarding the enhancement of the existing stack and the requirement for new functionalities. In this framework, solutions for re-transmission handling in multi-service, multi-AIV and multi-layer NR are further discussed

One key aspect of the UP design is the RAN/CN interface considerations and the interactions with CN given different protocols and QoS models. This was explicitly discussed in Section 3.3 by providing a new QoS model for the flow to bearer mapping, comparison of the thin-pipe with the fat-pipe approach and an extensive comparison of protocol options for the UP RAN/CN Interface.

4 User Plane design considerations on overall RAN architecture

This chapter describes the architectural aspects of a 5G user plane design for the RAN. Figure 4-1 shows an overview of the most important considerations in this respect. It is based on a separation of the network functions in two dimensions: In a first split it is considered that parts of the processing takes place at a central unit (CU) and others at distributed units (DUs). More details on this are provided in Section 4.1. In the direction of a second split it is then investigated how CP and UP can be separated, including the required interfaces, which is the topic of Section 4.2. In addition, Section 4.3 covers the topic of Network Slicing and its influence on the user plane design.

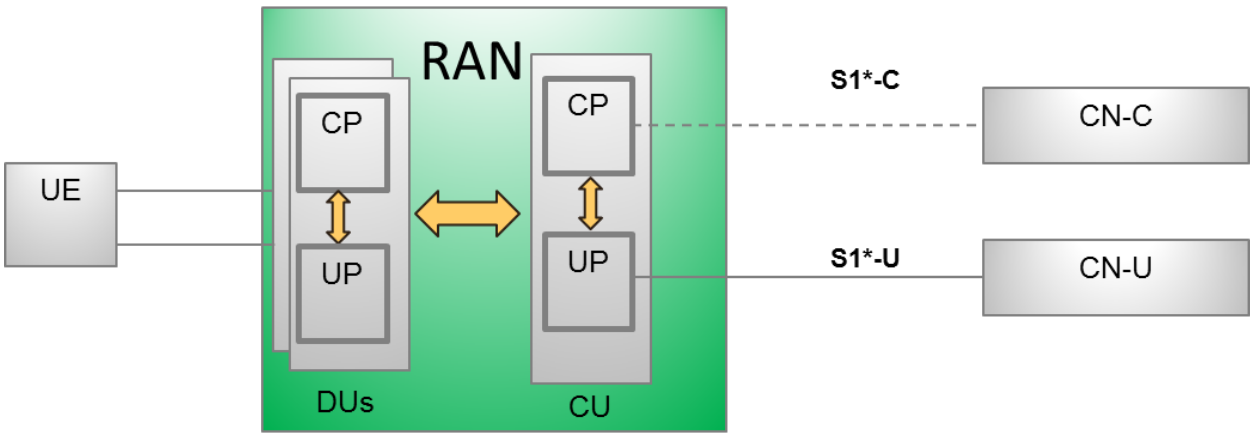


Figure 4-1: High level view on architectural aspects

4.1 Split of RAN user plane functions in central and distributed units

Different physical deployments and the capabilities of access nodes to support different AIVs will strongly impact the requirement and the efficiency of the UP Design. The designs presented in Chapter 3 were, to a large extent, deployment agnostic; here we look at practical limitations of different deployments and reassess Chapter 3 UP design solutions in this light. This first covers a functional split between a CU a DU for a single AIV in Section 4.1.1. Section 4.1.2 then describes the interworking with an evolved version of LTE and Section 4.1.3 introduces a service-based functional split.

4.1.1 Functional split within a 5G air interface variant

Cloud RAN (or Centralized RAN, C-RAN) [CCY+14] is an important technology that operates on top of Distributed Units (DUs) to benefit from resource pooling and centralized processing gains (hosted by a Central Unit, or CU) to handle inter-cell interference, to allow for higher resource utilization and to avoid peak-provisioning due to the centralized processing. In this context, multiple local C-RAN clusters can be formed to provide enhanced capacity in ultra-dense scenarios. The main drawback of this technology is the requirement for high capacity and ultra-low-latency fronthaul, which can be achieved mainly using ideal backhaul (fibre) for the Remote Radio Head (RRH) connectivity.

It is envisaged in a C-RAN scenario that the BS will be split into a CU and a DU to achieve a maximum network and radio resource efficiency by means of, e.g., centralized joint transmission, centralized scheduling, centralized flow control, etc. Section 5.5.2 of [MET216-D22] specifies all the possible functional split options between CU and DU, based on the 5G UP protocol stack. 3GPP also specifies similar options based on the LTE protocol stack [3GPP16-38801].

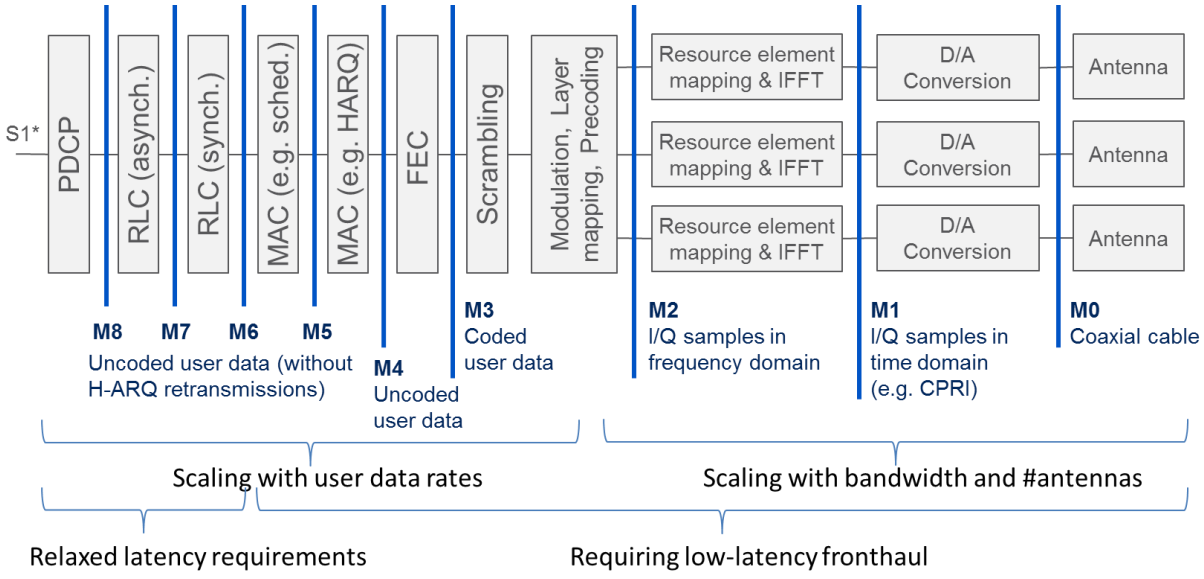


Figure 4-2: Function split options possibly implemented within the UP protocol stack [MET216-D22]

In a 5G RAN the UP protocol stack can be split at nine possible points between layers, designated as from M0 to M8, as shown in Figure 4-2. Those options put different data rate and latency requirements on fronthaul interfaces between CU and DU. The required data rate scales monotonically increasingly from higher layer splitting options to lower ones, notably dramatically at M1. This is depicted by an example in Figure 4-3, where cases with 2 to 8 antenna ports and different values for spatial layers are studied under the assumptions listed in Table C-6-1 (Annex C). Further calculations for different multi-antenna cases are shown in Figure 4-4 under

the assumptions provided in Table C-6-2, where 8 (LTE), 32 (5G AIV < 6 GHz), and 64 (5G AIV > 6 GHz) transmit antenna ports are studied, considering also suitable adaptation of OFDM parameter settings for the two 5G AIVs. The nearly unachievable extreme high data rates make option M1 infeasible especially in multi-antenna deployment scenarios. Based on the analysis, each split option has its own benefits and issues that need to be taken in deployment consideration. It should depend on the deployment scenarios which option to choose to form an optimal physical RAN structure.

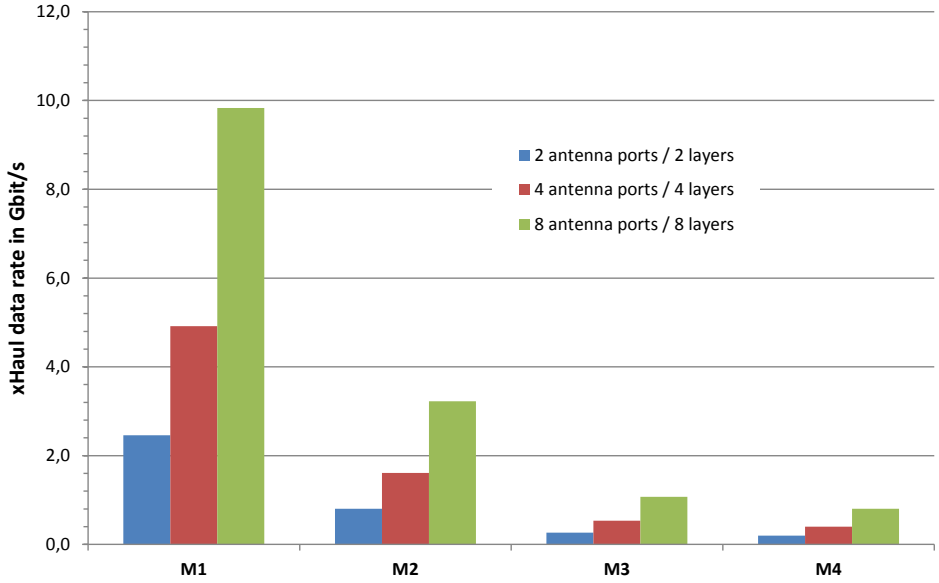


Figure 4-3: Downlink data rate for horizontal of CU/DU splits M1 – M4 considering LTE-like parameters

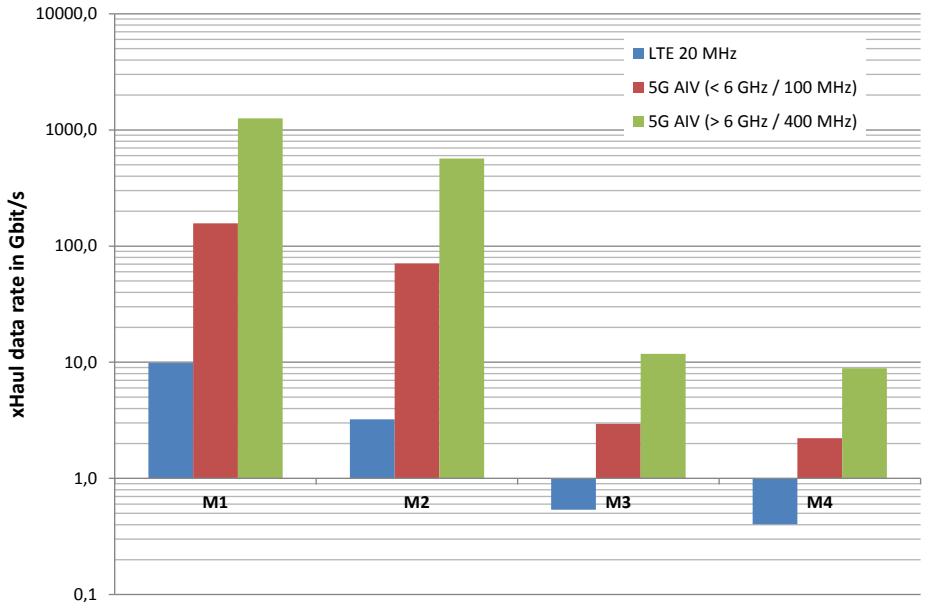


Figure 4-4: Downlink data rate needs of CU/DU split options M1 – M4, more multi-antenna cases

A functional split also impacts the small cells design greatly with particular applicability to urban ultra-dense environments. In such networks, numerous APs with various processing capabilities (from L1 to L3 APs) are deployed under the macro-cell (or a local cloud) umbrella [PMD+15]. Here, wireless in-band or out-of-band backhauling can be used to lower the Capital Expenditure (CAPEX)/Operational Expenditure (OPEX) costs of APs in urban environments. In this context, a parameter which can strongly affect the UP design is the availability, type and capabilities of wireless backhauling between APs (or DUs) and the CU. There can be different options for backhaul technologies (sub-6GHZ and mmWave above 6GHz) and topologies (point-to-point, point-to-multipoint, mesh) that may provide additional constraints (e.g. latency and capacity) which can strongly affect the functional operation and placement of a subset of protocol functions in the DUs.

As presented in [PMD+15], the backhaul (BH) capacity mainly depends on the BH technology and the topology used for the exchange between access points. In particular, the different types of BH topologies (point-to-point, point-to-multipoint, mesh) which can be deployed between a central unit (or macro or Gateway (GW)) and a multitude of access points, may strongly affect the latency and capacity of backhaul. Using the classification provided in [PMD+15], the fibre backhaul technology can be considered either as ideal backhaul (up to 10Gbps capacity), or as non-ideal (up to 1 Gbps). Regarding the wireless BH options, three key candidates are proposed for 5G RAN (sub-6GHz, microwave and millimetre wave (mmWave)). Sub-6GHz can provide a feasible solution for urban outdoor unplanned deployments where Line of Sight (LoS) is not possible and provides lower capacity (<100Mbps). On the other hand microwave and mmWave solutions can offer high capacity with proper planning (up to 2 Gbps).

In Table 4-1, the key features of the discussed candidate backhaul technologies can be summarized based on the classification described in [PMD+15]. In this table, the second column describes the end-to-end latency (one way from CU to DU) and the third column presents the throughput per hop latency (in case of multi-hop backhaul).

Table 4-1: Backhaul Classification [PMD+15]

BH technology	Total Latency (one-way)	Throughput
Ideal fiber access	2.5 µs	10 Gbps
Fiber Access 1	10 – 30 ms	10 Mbps – 10 Gbps
Fiber Access 2	5 – 10 ms	100 Mbps – 1 Gbps
Sub-6 GHz Wireless	5 – 10 ms	50 Mbps – 1Gbps
Microwave	< 1 ms	100 Mbps – 1Gbps
mmWave radio	< 1 ms	500 Mbps – 2Gbps

4.1.2 Functional split considerations including eLTE

If the operator has a legacy network deployed, there may also be enhanced LTE (eLTE) UEs connected the 5G Core Network serving legacy eLTE UEs in an overlapped NR coverage via an eNB. Non-standalone deployments based on eLTE+NR aggregation typically have two scenarios:

- A deployment where the enhanced LTE (eLTE) Radio Access Technology (RAT) is acting as the Master eNodeB (MeNB) and is providing at least the Primary Cell (PCell, typically on a lower carrier frequency). In addition one or more NR Secondary Cells (SCells) in the Secondary eNodeB (SeNB) can be configured (typically on higher carrier frequencies). This is occasionally referred to as “lean on eLTE”.
- A deployment where the NR RAT is acting as the MeNB and is providing at least the PCell (e.g. on carrier frequency F1), and in addition an eLTE SCell in the SeNB can be configured (e.g. on a carrier frequency F2) where both F1 and F2 are from a lower frequency band. This is occasionally referred to as “lean on NR”.

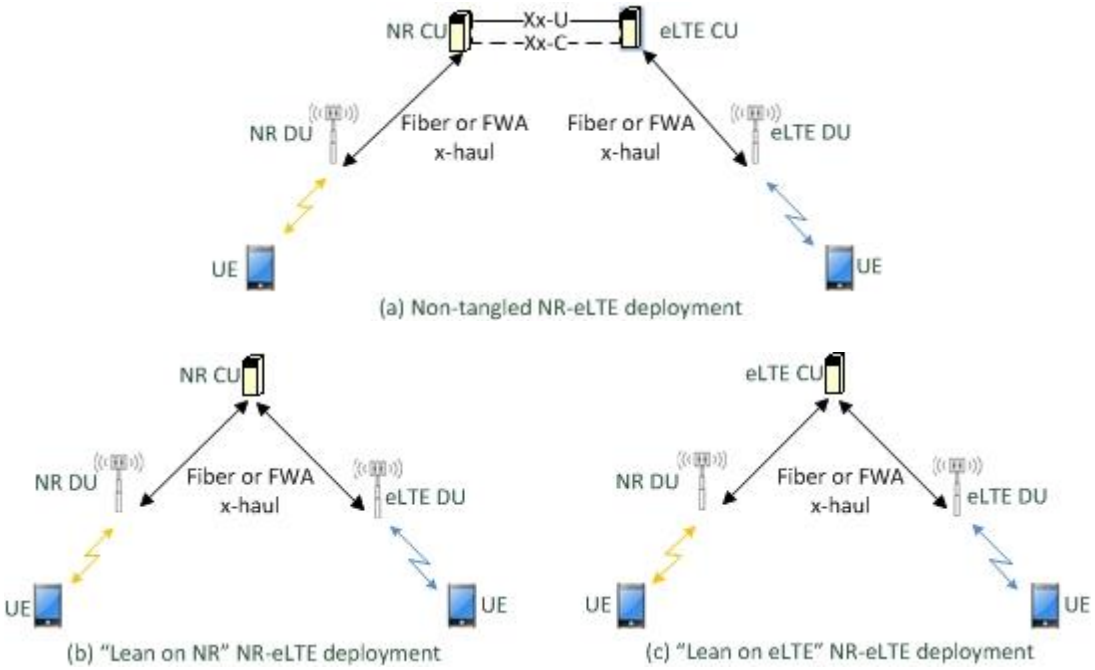


Figure 4-5: NR C-RAN and eLTE C-RAN deployment options with CU-DU splits

A particular case, which can apply mainly to a distributed RAN (D-RAN, without a CU), is the case of having access nodes which support a sub-set of AIVs. Two exemplary physical deployments are the following:

- DC to sub-6GHz NR and mmWave NR: In this deployment scenario, different nodes support different 5G AIVs at low frequencies and mmWave NR respectively. Sub-6GHz

AIVs are required to provide coverage, whereas the mmWave BS provides enhanced capacity in hotspot scenarios.

- eLTE-assisted NR: In this non-standalone NR deployment: eLTE BSs can support NR operation in high or low frequencies to further enhance capacity and coverage.

Section 2.3 of [MET217-D62] describes more in details on eLTE-NR tight interworking with regard to different AIVs.

The non-standalone scenarios do not impact much the CU-DU functional split schemes. In the New RAN there will be NR CUs and NR DUs in physical deployments, covering applicable radio frequencies in both higher bands and lower bands. There is no obligation that the NR design should be backward compatible. However, in an evolving co-existence scenario where a 5G RAN is collocated with an LTE RAN where LTE CU and DU are split, there could be different deployment options of CU and DU. The candidate deployments of NR-eLTE RAN with respective CU-DU functional split are shown in Figure 4-5. It is possible to connect eLTE CUs to NR CUs through the interface denoted as Xx in Figure 4-5.

4.1.3 Service-based Functional Split / Placement Considerations

To meet the diverse set of services (xMBB, uMTC, mMTC) requirements expected for 5G, one of the key enablers is utilizing the different AIVs. This will require the efficient aggregation of these AIVs. While considering different aggregation points in the protocol stack the impact needs to be analysed in the context with different architectural variants (centralized, distributed) and configurability of the protocol layers. All the functions provided by different protocols are not applicable to all services and might add to the complexity and overhead affecting the overall performance.

Recently 3GPP Release 12 has developed 4G DC solutions wherein a user equipment can receive data from two different transmission points called as Master eNB and secondary eNB [3GPP13-36842]. DC solutions aim to consider non-ideal X2 interface backhaul between the cells and aggregation over different carrier frequencies. The main advantage of DC is to improve robustness (mobility robustness) because of the existence of multiple data paths to the user [JSV+14, RPW+16]. DC solutions [3GPP13-36842] can be realized by either splitting the data in the core network, or by splitting the data at the MeNB.

As compared to 4G DC, 5G systems may utilize wider range of frequency bands i.e. both high frequency bands and low frequency bands, and more number of connected cells. Thus, 5G systems can utilize multi-connectivity solutions by using multiple cell groups and multiple bands at the same time. Multi-connectivity can thus be a rich source of transmit diversity in the network, which could be immensely beneficial for use cases such as ultra-reliable communications. One important aspect is how the user plane design can incorporate multi-connectivity solutions for simultaneously supporting multiple services.

The existing legacy system does not face the challenge of aggregating extremely diverse service requirements and of aggregating diverse AIVs to be supported simultaneously. In 4G and LTE, aggregation is possible at MAC layer with CA and at PDCP layer with DC.

The main contribution is to propose multi-connectivity options for support of multiple services. In this context, the following RAN design aspects have to be taken into account

- The support of multiple services at higher layers and lower layers
- Functional split options for DC/multi-connectivity.

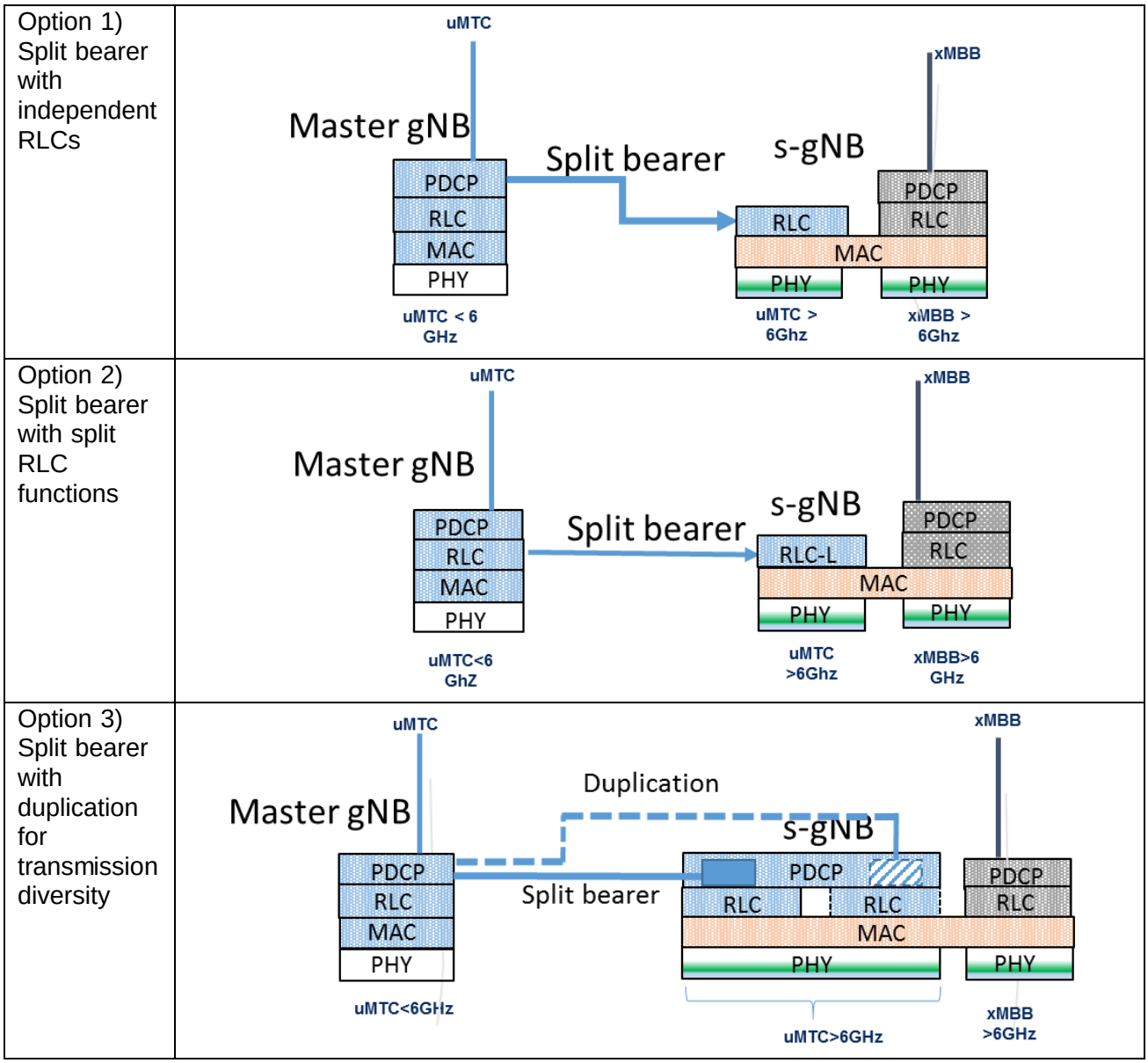


Figure 4-6: Multi-connectivity and aggregation of multiple services

The example below enumerates viable multi-connectivity options for supporting ultra-reliable communications (uMTC) and mobile broadband. Figure 4-6 shows an example of ultra-reliable communication uMTC service being supported using multi-connectivity from a master gNB and a secondary gNB, while enhanced mobile broadband is solely supported by a secondary gNB.

In Figure 4-6, options 1) and 2) show two examples of applying 4G split bearer solution (mobile broadband) for 5G uMTC while also aggregating with mobile broadband services. The 4G split bearer DC solution effectively splits the packets at the master node, which means that part of the packets are sent to the secondary node for transmission to the user terminal. In option 1) the packets are separated at the PDCP layer, while in option 2) the packets are separated at the Radio Link Control (RLC) layer. In option 1) all the RLC functions such as buffering, ARQ, concatenation/ segmentation are implemented at the secondary node for the spit bearer. However in option 2) the RLC functions are split between the master and secondary, i.e., some of the above RLC functions are handled by the master node, while some other functions could be handled at the secondary. The solution in option 3) specifically targets to make use of transmission diversity across different cells for 5G uMTC while also providing service-specific configurability. Note that support of massive machine type communication is not shown for the sake of convenience.

The design approach is described below.

- Service aggregation: The different services (i.e. xMBB and uMTC) are shown in different colours (i.e. grey and blue). In order to ably support multiple services while also keeping service specific functionalities, and simplify handling of mobility robustness during handover, the flows are multiplexed at the higher layers. This means that different PDCP and RLC entities are enabled for each of the services.
- Harmonization of protocol functions: The protocol functions (PDCP, RLC, MAC) are harmonized across different services and different transmission legs, shown as dotted pattern. This means that the protocol functions (e.g. handling of RLC transparent mode, acknowledged mode, unacknowledged mode (UM)) are commonly defined while different functional instances may be used for different services. In such a case, while one uMTC service flow can be configured with RLC AM, an xMBB service flow may configured with RLC UM thus allowing for cross-optimization between the flows.
- One key aspect is to utilize multiple transmission points using multi-connectivity. In principle, multi-connectivity improves robustness of data reception. As it can be seen in the examples above, multi-connectivity refers to three different flows, uMTC from the master gNB, uMTC from the secondary gNB (s-gNB). Such an example can be extrapolated to more than 2 cells, for example realizing uMTC from another s-gNB. For the purpose of multi-connectivity, different PDCP, RLC and MAC entities are established per transmission leg. However, to enable robust support of multiple services, different PDCP and RLC entities are also established per service flow.
- Different service types are handled by one MAC entity per transmission leg. Thus service specific configurations can be optimized via service specific functional

configurations within the MAC entity. On the other hand, different MAC instances may be realized in different transmission legs using the same harmonized MAC functions.

- In the case of option 3), transmission diversity is further enhanced by duplicating the master gNB's data at the secondary gNB. The duplicated data can then be used to enhance diversity whenever required by a specific service type. Thus whenever required a user terminal may benefit from diversity by receiving data of a service flow from both master gNB and the secondary gNB.
- The master gNB and the s-gNB may be deployed in different frequency bands, e.g. below 6GHz and above 6GHz. The utilization of a harmonized PHY layer is shown as shade of green and white. This indicates the PHY layer reuses harmonized functionalities and waveforms between the lower (white) and the higher (green) frequency bands. It also means that the same harmonized PHY function is then used for both xMBB and uMTC in the s-gNB. As an example, the higher frequency bands may reuse the waveform functions from the lower frequency bands, while in addition realizing band specific functions such as beamforming in the higher bands. Service-specific parametrization can then be made within the harmonized PHY, for example different service types may utilize different beamforming configurations at the higher bands.

5G multi-connectivity can thus build on 4G DC solutions, while also utilizing new features such as data duplication for diversity for 5G services. 5G multi-connectivity solutions should also strive to effectively aggregate multiple services while maintaining service-specific configurability.

4.2 Separation of RAN control plane functions from user plane functions

This section describes the concept of a separation of the CP and the UP in the network architecture. In the following the motivation for such a split is discussed and a network architecture that implements a full split is proposed. As a full split might be complex and difficult to implement, also partial splits are introduced.

In addition to what is presented here, Section 2.2 of [MII17-D62] describes potential deployment architectures for CP and UP functions.

4.2.1 Motivation

The network functions (NFs) of a wireless network are typically categorized into two groups: The UP is responsible for forwarding data from the source to the destination, including the corresponding processing. The CP controls the UP, for example in terms of setting the routing path of a packet or of radio resource management. The CP also provides a set of other functionalities such as connection / mobility management and broadcasting of system information.

The separation of CP and UP according to the SDN concept is a recent trend in the definition of the 5G architecture [TGV+14, RBB+16, 5GPPP16]. It requires categorizing all NFs as being

either part of CP or UP based on functional decomposition [5GPPP16, MET216-D22]. Any kind of interaction between CP and UP is supposed to happen through standardized interfaces. For the CN a separation of CP and UP functions is already partially realized e.g. in LTE with the Mobility Management Entity as the main CP element and the Serving and Packet Gateways as UP elements. However, this separation is not complete, e.g. the Packet Gateway contains CP functions. An enhanced separation of CP and UP for the ECP is under study in [3GPP16-23214].

For the RAN the anticipated benefits of a CP/UP split are:

- In multivendor networks, a standardized interface to the CP enables a consistent control over network elements and NFs from different vendors / manufacturers, e.g. in terms of interference management for ultra-dense networks [MET216-D22, MET216-D51].
- Due to the tight coupling of CP and UP NFs in today's networks, the replacement or upgrade of a CP function often requires also the replacement of UP functions. Avoiding this might offer significant cost savings.
- The independent evolution of CP and UP by possibly modifying and adding CP functions without changing the UP (and vice versa) could make the rollout of new NFs faster thus enable a more flexible network.

Besides, there are also disadvantages:

- CP and UP functions are often tightly coupled, especially in the lower radio protocol stack layers. It might be challenging and could affect the performance when fully separating CP and UP handling, especially if the processing is not collocated.
- Standardization is required in case the interfaces between CP and UP have to be extended to introduce new features which might slow down this process. Integrating additional interfaces in a proprietary manner in combination with standardized ones is not a suitable solution, as it would destroy the benefits of a CP/UP split. For example, a flexible change of CP NFs in logical network elements would not be possible any more if only selected UP NFs support certain proprietary interfaces.
- Additional effort in terms of testing is required to guarantee the interoperability of CP and UP functions from different sources (shifting the effort to system integrators supporting the operators instead of doing this work at a single vendor).

4.2.2 CP/UP split based network architecture

In this section a RAN design concept with a full CP/UP split between is described. It covers the transport as well as the access network and uses also a split into CUs and DUs.

Figure 4-7 gives an overview of the proposed architecture. The CUs in the RAN are the Central Access Controllers (CACs) that centrally host CP and UP functions. They are split into an UP part (CAC-U) and a CP part (CAC-C). Typically, the lower layers of the radio protocol stack are hosted close to the antenna sites, whereas the higher layers are processed at the CAC, but in principle also a fully flexible allocation is feasible.

The transport network (aggregation) which forwards the UP data from and to the CN is implemented through SDN switches or routers. With respect to traffic routing the CN mobility management function [3GPP16-23799] acts as the responsible SDN controller. The main role of the SDN controller is to enforce that data is forwarded to the correct antenna site, especially in case of mobile users.

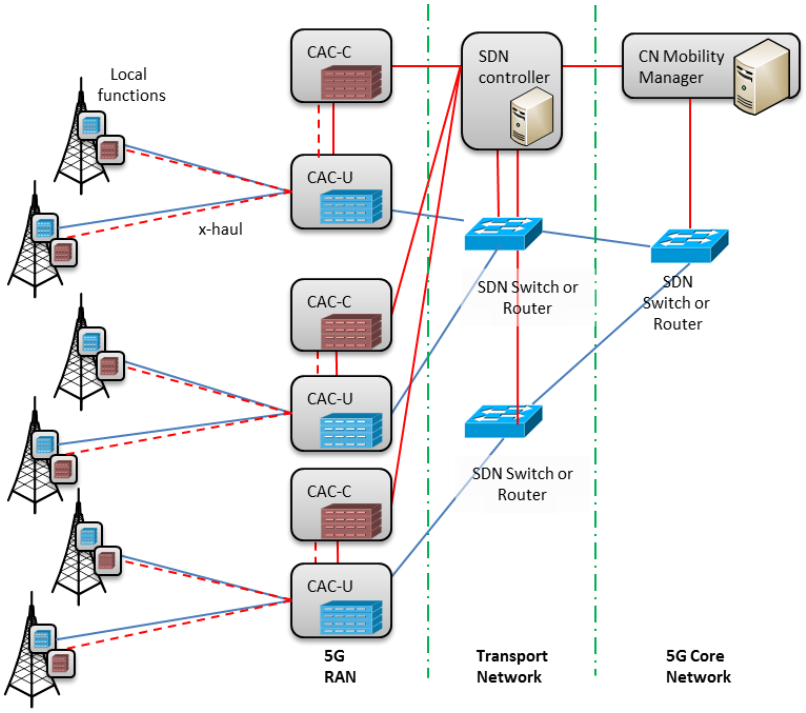


Figure 4-7: SDN-based 5G network architecture supporting flexible functional CP/UP splits especially in the radio access (not all CN functions are shown)

Beside the already mentioned general advantages of a CP/UP split, this SDN-based approach offers additional improvements compared to legacy tunnel-based approaches as the GTP based solution in LTE and UMTS, such as reduced overhead and improved integration with fixed networks as described in Section 3.3 and in [YHZ+16].

To realize a scalable approach it does not make sense to implement a country-wide RAN via a single CAC (or CU, respectively), but to implement several CACs each controlling the radio processing for a certain number of antenna sites (domain). Suitable locations for CACs are e.g. the central offices of fixed or integrated network operators [5GPPP16]. To support especially low latency applications, mobile edge computing facilities [RBB+16, 5GPPP16] can be integrated into the CU. Typically, the NFs running in the CU (CAC-C/U) are implemented as virtual functions on server platforms based on Network Function Virtualization (NFV) principles [ETSI-NFV].

In the presented architectural approach, three cases of user mobility handling are possible:

1. Between the sites within the domain of a CAC, mobility is handled CAC-internally. This can happen through fast UP switching [MET216-D51]. In that case no signalling traffic is required between RAN and CN. More details on mobility inside a central unit can be found in Section 6.2 of [MET217-D62].
2. Inter-CAC-U handover: Here the user equipment (UE) moves from one CAC domain into another one. If both CACs are connected to the same SDN switch or router the SDN controller of the transport network can simply trigger the redirection of the data flow.
3. CN-based handover: In case a path switch has to happen at the highest level (CN-based), it is under the responsibility of the CN mobility manager to send a command to the SDN switches/routers of the CN. In addition, the new route in the transport network has to be set by the corresponding SDN controller.

The cases 1 and 2 describe a RAN-based mobility, where the mobility handling happens only within the RAN. This is beneficial because of low latency between involved components and therefore a low handover interruption time (ideally zero). This advantage is especially relevant for ultra-dense radio node deployments (using e.g. mmWave bands) with a high number of mobility events [MET216-D51].

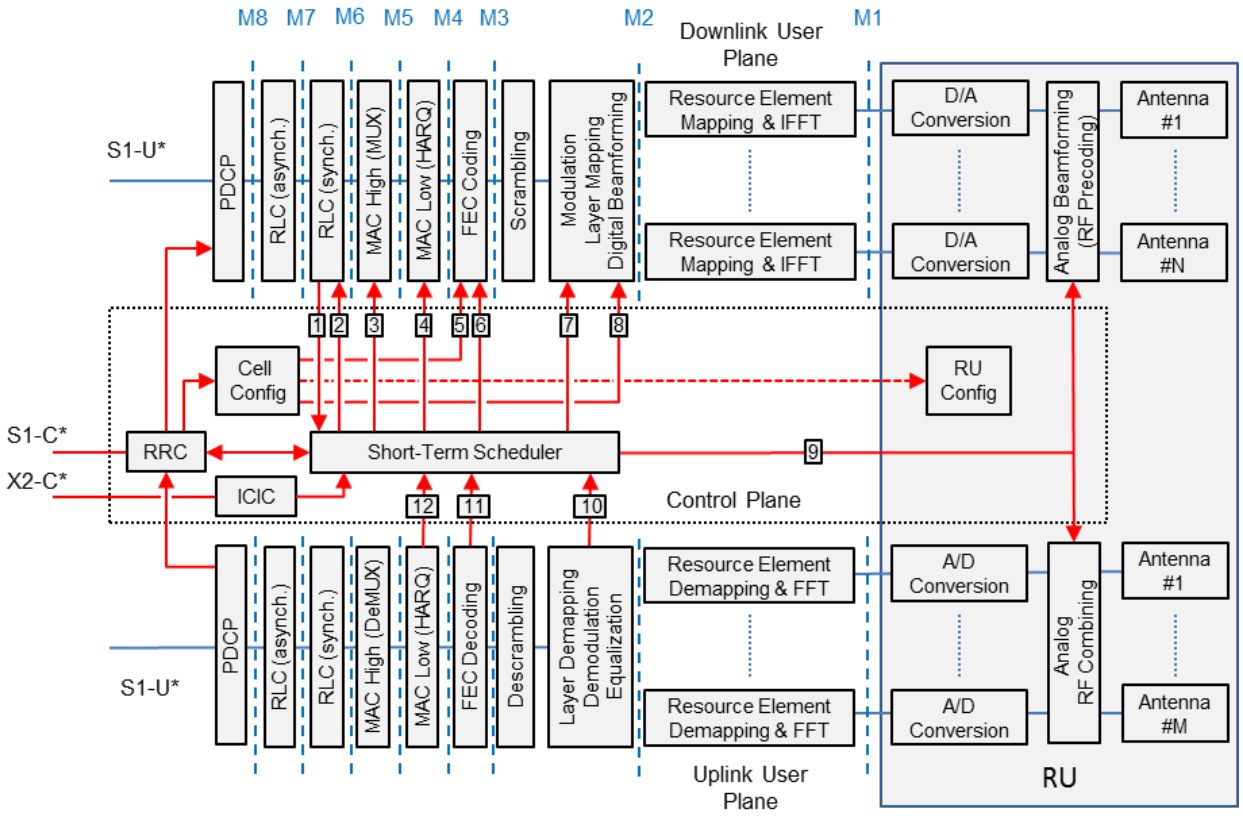


Figure 4-8: Control and user plane decomposition and interactions in the radio access network (network infrastructure part only; single radio protocol stack)

With respect to the processing in the RAN a more complex interaction between CP and UP is required in case of a CP/UP split. Figure 4-8 shows the UP processing chain for downlink (DL; upper part of the figure) and uplink (UL; lower part). The CP functions are separated in the middle of the figure. The interactions between CP and UP are indicated by arrows and described in the following (please note: only main interactions are shown to not complicate the figure).

The CP NF RRC implements the corresponding 3GPP protocol layer. It is mainly responsible for the establishment, maintenance and release of connections to the UEs. The required interaction with the UEs happens by generating RRC control messages, which are then forwarded to the UP. By handing over the generated messages to the PDCP layer, they enter the UP processing chain and are finally transmitted through the antennas. Corresponding RRC messages generated by the UEs are processed by the UL UP chain and then forwarded to the CP NF. Thus a full communication between the CP NF RRC and the UEs is enabled through the UP.

The CP NF "Cell Configuration" is responsible for transmitting cell information (e.g. the cell identification) and setting basic cell parameters (e.g. transmit power and electrical tilt).

This happens via sending broadcast information and reference symbols through the UP (interactions 5 and 8) and by configuring the Radio Unit (RU).

The scheduler represents the CP NF with the strongest coupling to the UP. The following interactions with the UP have been identified and are indicated with corresponding numbers in Figure 4-8:

1. DL buffer status: DL data arrives from the CN through the S1-U* interface (via the transport network). It is processed by PDCP and RLC layer which then reported to the scheduler that data for DL transmission is available.
2. Payload selection: The scheduler selects data to be forwarded to the MAC layer.
3. DL resource assignment and generation of UL transmission grants: In the DL, this enables the MAC layer to generate corresponding transport blocks. For the UL transmission grants are generated and transported by the UP to the UEs.
4. Retransmission control: Retransmissions by means of HARQ are also controlled by the scheduler, who sends the corresponding commands to the UP.
5. Coding scheme: The scheduler sets the coding rate to be applied (per UE) and configures the UP accordingly.
6. Antenna mapping, precoder, modulation scheme: Similar to coding scheme, the scheduler also configures the modulation scheme to be applied. For MIMO operation, also antenna mappings and precoder settings are required at the UP.
7. In case of analog beamforming (e.g. for Massive MIMO), the scheduler sets the corresponding antenna weights used in the UP.
8. CSI from UL sounding: In UL, after demodulation, CSI can be generated based on sounding sequences that the UEs sent.
9. CSI from reporting, UL scheduling request: After the demodulation the CSI information from reporting is available. Also scheduling requests for future UL transmissions have to be forwarded to the scheduler.

10. HARQ status: The scheduler receives the status of UL and DL HARQ processes, e.g. acknowledgements.

The ICIC also acts as CP NF, but works in contrast to the scheduler on a long-term basis, i.e. not on TTI level.

Figure 4-8 visualizes the tight coupling of CP and UP in the RAN, which also shows the huge effort in terms of standardization that is required to achieve a full separation.

4.2.3 CP/UP-split variations in RAN deployments

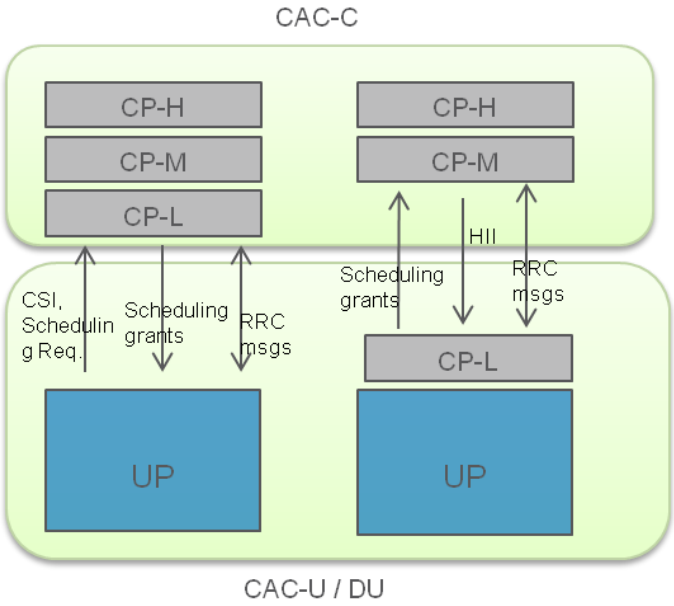


Figure 4-9: CP/UP Split Options in the 5G RAN

In RAN deployments, due to the strong coupling of UP with control logics, the complete CP/UP separation, might not be feasible for all different deployments. In particular, there might be real-time Scheduling (or Control Plane-Low (CP-L) as discussed in [MET216-D62]) functionalities which will require per TTI scheduling. In case of a D-RAN, the de-coupling of real-time RRM will require ideal BH / Fronthaul between the control part of gNB (CAC-C as mentioned above) and the data processing part of gNB (which can reside either at CAC-U or DU). Hence, below we show some cases where we partially split some C-Plane functionalities. Since there might be multiple interactions between independent functionalities, we group them in two types of C-Plane functionalities: 1) CP-L or Real-time Scheduling, 2) CP-M/H or Non-Real time Scheduling (Cell Re-Selection, Connection mobility control, Load Balancing, ICIC) and RRC. In Figure 4-9, we show different options and the pros/cons of applying them to different deployments:

Option 1: Complete Separation of C-U Plane: Only control logics remain at CAC-U (ARQ, HARQ)

- Pros: High gains due to Centralized Scheduling, virtual /No-cell concept can be realized with this split.
- Cons: Ideal BH required

Option 2: Separation of CP-H, CP-M: Real-time RRM remain at CAC-U (DRA/PS)

- Pros: good compromise for D-RAN with non-ideal BH
- Cons: dependencies between RRM functions will require additional signalling (e.g. High Interference Indicator (HII) might be required to be exchanged between CAC-C and CAC-U)

4.3 Impact of Network Slicing on the User Plane

4.3.1 Network slicing in a 5G RAN

A Network Slice is a virtual network created by the network operator customized to provide an optimized solution for a specific market scenario which demands specific requirements with end to end scope as described in [3GPP16-23799] [3GPP16-38801]. Basically Network Slicing is introduced to provide additional means to ease the service management and provision in 5G networks. It is envisaged that Network Slicing will primarily be business driven, where each Slice will support one or more 5G services [MMM16-W31]. Network Slicing extends the comparatively static principle of network sharing [3GPP14-22852]. It is described by NGMN to include the RAN while a Network Slice is defined as an end-to-end network [NGMN15]. NGMN specifies the Networking Slicing concept consisting of 3 layers [NGMN16]:

- Service Instance Layer
- Slice Instance Layer
- Resource Layer including radio resource

Each Network Slice should be capable of being accessed by the UE and managed by the Slice owner as a logically independent network. As different parallel Network Slices may target a variety of use cases with very diverging QoS requirements, there are a few principles that can be made on the RAN design, in relation to the UP design:

- Visibility of Slices to the 5G RAN is required to be able to apply Slice differentiation
- Sharing of most RAN resources between multiple Slices is assumed as default
- 5G RAN should support differentiation of traffic between Slices
- 5G RAN should support QoS differentiation within a Slice

3GPP specifies the solutions for supporting Network Slicing in [3GPP16-23799], where the Solution 1.1 introduces the high-level approach without slicing the radio. It is reasonable because high level Slices should be independent of radio features though they will utilize radio

resources. The RAN Slicing is being discussed in 3GPP [3GPP17-R21701510] and it is envisioned the RAN Slicing will work together with CN Slicing, while with a major impact on RAN CP rather than RAN UP, e.g. on mobility management, session and connection management, and network resource management etc. The aspects on the CP are comprehensively addressed in [MET216-D51] and [MET217-D52].

With respect to the UP design, the main challenge is to simultaneously address the potentially diverging requirements of the different Slices, e.g. in terms of latency, data rate and reliability. Due to the fact that an efficient multiplexing of different services (xMBB, uRLLC and mMTC) is a requirement for 5G since the beginning, this aspect can be addressed in the UP design in the following ways:

- Due to multi-waveform harmonization, multiple AIVs with different characteristics can be combined efficiently (see Section 2.2)
- Different numerologies can co-exist within a single frequency band (see Section 2.4).
- The protocol configuration is designed for multi-service support (see Chapter 3).

4.3.2 Network Slicing impact on CU-DU functional split

Network Slicing won't impact much on the UP of a D-RAN. The Slicing related procedure, e.g. Network Slice Instance selection and cell selection, can be executed in the BS with a reasonable performance. In a C-RAN scenario, Network Slicing may impact the CU-DU functional split granularity. Apart from a functional split per CU, DU, UE or bearer, there is an option for a functional split per Slice, provided that it can be justified. The option may come from a finer granularity demand. For this option it is expected that each Slice would have at least some distinctive QoS requirements. Regardless of how exactly a Slice is implemented within the RAN, different functionality mapping may be suitable for each Slice. The option implies that a particular instance of the interface between a CU and a DU would need to support simultaneously multiple granularity levels on the UP. Nevertheless, the option is not a baseline option for a CU-DU functional split. It will impact the baseline CU-DU functional split options preferences, based on QoS and latency requirements though.

5G networks will rely on NFV to realize end to end Slices. It is envisaged that the 5G RAN will be implemented as part of a NFV Infrastructure and support Virtualized Network Functions. That will impact the BS implementation and the CU-DU functional design in the C-RAN scenarios. Supporting diverse services required for, e.g. xMBB, uRLLC, mMTC, will make some of the CU-DU functional split options more preferable. If the split is realized at a higher layer of the protocol stack, more flexibility is allowed in the deployment of DUs as there are less constraints on latency and data rates. This is reflected by a simple comparison in Figure 4-10. Some Slicing related signalling (Slice selection as part of a Common Control Network Function, etc.) may be handled within the DU if the higher layer functions are available at the DU and allow the DU to deal with the UE signalling immediately. That would save the x-haul transportation to the CU

otherwise. Referring back to results summarized in Section 4.1.1, that makes options M8, M7 and M6 described in Section 4.1.1 desirable as they provide more flexibility in the RAN planning and accommodating QoS requirements alongside with Network Slicing. This gain could be minor if a high quality x-haul is in place, though. Based on the analysis options M8, M7 and M6 should be considered first. A preference on higher layer split option is also the trend in eLTE. The similar preferable 3GPP options 2 and 3 are also being studied further by 3GPP [3GPP17-R21700637], where option 2 is comparable to M8 and option 3 to M6.

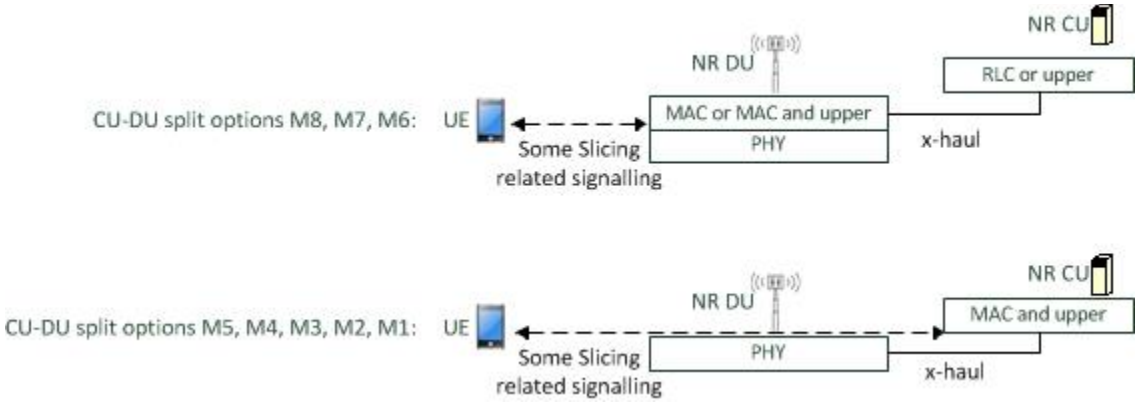


Figure 4-10: A simple comparison of different CU-DU split options regarding RAN Slicing

4.4 Conclusions on user plane design considerations on overall RAN architecture

This chapter described the architectural aspects of the user plane of a 5G Radio Access Network. It covered the split of network functions in two domains: Functions can be separated according to where the processing takes place (in a central unit or a distributed unit) and according to their purpose (processing user plane data or controlling other functions).

A flexible functional split enables that for different physical infrastructure centralization can be realized such that the corresponding gains (e.g. from coordination or traffic steering) can be exploited. However, it was also shown that functional splits at a lower protocol layer can be extremely demanding for the interface between central and distributed unit, especially in terms of data rate for a high number of antennas.

With respect to a split of functions into the control plane and the user plane functions, it was shown how an architecture based on a full split could be realized, including the corresponding interfaces. A standardization of all required interfaces seems complex, such that a partial split rather than a full separation is more realistic.

A 5G RAN can utilize new features in terms of multi-connectivity such as data duplication for diversity for 5G services. 5G multi-connectivity solutions should also strive to effectively aggregate multiple services while maintaining service-specific configurability.

The 5G RAN will provide support for Network Slicing based on Network Function Virtualization. Network Slicing itself does not impact much on UP in 5G RAN as the resource is shared by default and RAN Slicing could be addressed on CP. Network Slicing won't impact much on a D-RAN UP. However it may impact the CU-DU split options and make some of the higher layer split options like M8, M7 and M6 a bit more preferable.

5 Conclusions

In addition to the earlier METIS-II deliverable D4.1 [MET216-D41], this document contributed to the Holistic Air Interface Harmonization Framework which is one of the key innovation pillars in the METIS-II project, as part of the overall 5G RAN Design objective. In this closing section, we examine how the goals of this document, which is the final external deliverable from METIS-II WP4 were achieved. We also summarise what the key outcomes from the work reported in this study are, and how they may impact final work in METIS-II.

This document had the following three aspiring goals:

- **Objective 1**: to provide an analysis using the METIS-II AI design and evaluation framework of physical layer enablers (new waveforms, enhanced spectrum up to 100GHz, different frame structure and numerologies) to be introduced in the 5G context. Particular focus is given to multi-service coexistence considerations and harmonization, integration among AIVs and with legacy technologies, giving examples of various options and underlying trade-offs. Preliminary technical work on these aspects was extensively considered in chapter two, three and four of D4.1;
- **Objective 2**: to expand the analysis of a corresponding flexible 5G user plane design and harmonization with new protocol stack functions and mechanisms required to support diverse service requirements of the 5G service families, i.e., xMBB, mMTC, and uMTC. Deliverable D4.1 included an initial study of harmonized user plane design and also protocol aggregation approaches and this deliverable complements those considerations;
- **Objective 3**: to present different user plane architecture concepts including software defined architectures and study how various user plane design decisions, in particular related to the functional split options, a control / user plane split and Network Slicing impact the overall 5G architecture.

With a view to achieving these objectives, the following main contributions were proposed:

- **(Objective 1)** A harmonized and unified way of describing the 5G air interface design proposals using main system enablers was proposed, including a proposal of a harmonized implementation of multiple 5G envisioned waveforms, taking into account the performance analysis and spectrum considerations. In addition, from the final air interface harmonization and user plane design point of view, preliminary answers to some of the METIS-II key RAN design questions were also provided, in particular:
    - **AIVs expected to be introduced in 5G** – although 3GPP in 5G Phase 1 has agreed on one AIV, there will be future studies that may include other AIVs to support the heterogeneous demands of the 5G services. The final 5G AI may then, be composed of different AIVs. In particular, from user plane design perspective and for legacy applications, an AIV for lower spectrum bands (below 6 GHz) is expected to be the evolution of the current 3GPP LTE standards. Above 6 GHz, new AIVs with special frame structures may be required to better

support more advanced techniques (i.e. massive MIMO and channel estimation mechanisms). New applications, which could be characterized by a strong asynchronous effect (e.g. D2D, vehicular communications), may require special waveforms to efficiently counteract it. Furthermore, V2X road safety applications may require new AIVs to effectively deal with their multicasting nature and high reliability requirement.

- **AIVs that 5G devices should support** - this decision will depend on the purpose and the type of the device. The harmonized 5G AI should allow that purpose-specific devices implement only necessary modem functionalities. For instance, a laptop thought to be in a static or quasi-static indoor environment may implement AIVs for below and above 6 GHz with massive MIMO support and high-order modulations. On the other hand, an in-car communication unit may implement an AIV with high reliability and multicasting features suitable for V2X communications. Hence, this in-car unit may not necessarily require more advanced modem features like high-order modulations or massive MIMO support.

Furthermore we provided guidelines on a suitability of 5G spectrum bands for all three service types considering coverage and capacity trade-offs below and above 6 GHz including a discussion on higher frequency bands suitability for V2X and uMTC communications. To support conflicting service requirements in the same band, a more flexible AI approach would be preferable including different frame structures and numerologies while also a harmonized approach to band resource usage was studied. Finally, as part of the air interface considerations, by using a mathematical approach we also identified that coherent and non-coherent reception techniques could be harmonized in one receiver.

- **(Objective 2)** After the study of different PHY layer enablers was presented, an investigation of a harmonized 5G user plane design in higher layers (MAC, RLC, PDCP) was carried out with regards to a flexible service oriented set of features. As part of this analysis, an initial assessment of the legacy LTE protocol stack was performed, identifying major challenges and shortcomings. As a result, new protocol functionalities were proposed either as an extension to the existing functions or new approaches and solutions. In particular, improved 5G retransmission mechanisms in MAC/RLC suitable for addressing multi-service scenarios were proposed including their performance evaluation. In addition, user plane aspects were investigated as part of the improved 5G QoS model considerations and also a comparison of different protocol stack approaches to the interface between radio and core network including their trade-offs was discussed. This sets the scene for an overview of the current 5G UP standardization progress in 3GPP.
- **(Objective 3)** Finally, architectural aspects of the user plane of an envisioned 5G RAN were examined in more detail. New approaches to functional split based either on processing based (central vs distributed) and purpose based (user vs control) solutions were investigated and their performance trade-offs were presented. More flexible physical network infrastructure implementations and centralization gains (e.g. coordination, traffic engineering) could be achieved by introducing central and distributed nodes but taking into account higher interface requirements (e.g. latency), especially

when the functional split is placed lower in the protocol stack. Although it was identified that the full user plane / control plane split architecture was possible from the radio access network implementation point of view, potentially improving deployment flexibility for the operator, in practice it should be balanced with the additional cost of introducing new standardized interfaces between network nodes. Furthermore, as part of the considerations on network slicing impact on UP, we concluded that the effect on the user plane design should be less significant in comparison to the control plane especially when a distributed-RAN is deployed. However, it needs to be carefully considered especially when a central / distributed unit split is introduced.

A key outcome of the work presented in this deliverable is a further study and a detailed insight into the harmonized 5G air interface framework presented in Chapter 2 and user plane design presented in Chapter 3 and 4. In addition, this contribution comprehensively addresses initial METIS-II WP4 objectives (identification of most promising air interface candidates, their user plane harmonization and development of a common user plane design) including two main tasks assigned to WP4 (this is complementary to the 5G AI candidates harmonization work presented in METIS-II D4.1 [MET216-D41]):

- Task 4.1 - Air interface analysis and user plane harmonization;
- Task 4.2 - Architectural aspects of a common user plane framework for the 5G air interfaces.

Based on our work, we conclude that considering the broad landscape of new 5G system requirements, a fully harmonized air interface and user plane design framework may seem an ambitious target but we believe that the provided contributions make a leap step in this direction and provide useful guidelines in the design of a new system.

It is envisioned that the elaborated evaluation insights, resulting from a wide METIS-II consensus, aligned with 3GPP while offering a long-term, integrated system view, will impact researchers and standards bodies in different trade-offs they consider when assessing new approaches to AI and user plane design.

6 References

[3GPP13-36842] 3GPP TR 36.842 v 12.0.0, “Study on small cell enhancements for E-UTRA and E-UTRAN, High Layer Aspects”, December 2013.

[3GPP14-22852] 3GPP TR 22.852 V13.1.0 (2014-09), Technical Report, “3rd Generation Partnership Project; Technical Specification Group Services and System Aspects; Study on Radio Access Network (RAN) sharing enhancements (Release 13)”, September 2014.

[3GPP15-36300] 3GPP TS 36.300, Technical Specification Group Radio Access Network Evolved Universal Terrestrial Radio Access (E-UTRA) and Evolved Universal Terrestrial Radio Access Network (E-UTRAN), “Overall description,” Release 13, v13.2.0 (2015-12).

[3GPP15-36322] 3GPP TS 36.322. Technical specification group radio access network; Evolved Universal Terrestrial Radio Access (EUTRA); Radio Link Control (RLC) protocol specification, Release 12, v12.4.0 (2015-01).

[3GPP16-1612724] 3GPP RP-1612724 NTT DOCOMO, “Frequency-domain aspects of frame structure,” November 2016.

[3GPP16-166093] 3GPP R1-166093, “Waveform evaluation updates for case 1a and case 1b,” 2016

[3GPP16-167376] 3GPP RP-167376 NTT DOCOMO, “Comparison of candidate waveforms,” August 2016.

[3GPP16-23214] 3GPP TS 23.214, Architecture enhancements for control and user plane separation of EPC nodes, Release 14, v14.2.0, March 2017

[3GPP16-23799] 3GPP TR 23.799 V14.0.0 (2016-12), Technical Report, “3rd Generation Partnership Project; Technical Specification Group Services and System Aspects; Study on Architecture for Next Generation System (Release 14)”, December 2016.

[3GPP16-25912] 3GPP TR 25.912, Feasibility study for evolved Universal Terrestrial Radio Access (UTRA) and Universal Terrestrial Radio Access Network (UTRAN), V13.0.0 (2016-01).

[3GPP16-29281] 3GPP TS 29.281, General Packet Radio System (GPRS) Tunneling Protocol User Plane (GTPv1-U), Release 13, V13.2.0 (2016-06)

[3GPP16-38801] 3GPP TDoc RP-162255, TR 38.801 V1.0.0 (2016-12), Technical Report, “3rd Generation Partnership Project; Technical Specification Group Radio Access Network; Study on New Radio Access Technology; Radio Access Architecture and Interfaces (Release 14)”, December 2016.

[3GPP16-RAN1-86] 3GPP RAN1-86 Chairman Notes, September 2016.

[3GPP16-RAN1-86b] 3GPP RAN1-86bis Chairman Notes, November 2016.

[3GPP16-RP163476] 3GPP TDoc RP-163476. Beam Management in NR, Release 14.

[3GPP17-38801] 3GPP TR 38.801. Study on New Radio Access Technology; Radio Access Architecture and Interfaces" v1.1.0 (2017-03)

[3GPP17-38804] 3GPP TR 38.804. Study on New Radio Access Technology; Radio Interface Protocol Aspects V0.5.1 (2017-02)

[3GPP17-R21700637] 3GPP TDoc R2-1700637 "Summary of RAN3 status on CU-DU split Option 2 and Option 3, and questions/issues for RAN2", RAN3 NR Rapporteur (NTT DoCoMo, Inc.), January 2017.

[3GPP17-R21701510] 3GPP TDoc R2-1701510, "Mobility and RAN slicing", Sony, February 2017.

[3GPP17-TSG_WG2] Chairman notes, 3GPP TSG-RAN WG2 NR Ad Hoc, Spokane, USA, 17th – 19th January 2017.

[5GPPP16] 5G PPP Architecture Working Group: White Paper "View on 5G Architecture", Version 1.0, July 2016.

[ABS11] P. Achaichia, M. L. Bot, and P. Siohan, "Windowed OFDM versus OFDM/OQAM: A transmission capacity comparison in the HomePlug AV context," in IEEE International Symposium on Power Line Communications and Its Applications (ISPLC), 2011, pp. 405–410.

[BH03] B. Hassibi and B. M. Hochwald, "How much training is needed in multiple-antenna wireless links?," IEEE Trans. Inf. Theory, vol. 49, pp. 951–963, Apr. 2003.

[BKP+16] G .Berardinelli, S R. Khosravirad, K I. Pedersen, F .Frederiksen, P. Mogensen, "Enabling early HARQ feedback in 5G networks", VTC-Spring 2016

[BKW+17a] J. Bazzi, K. Kusume, P. Weitkemper, Kazuaki Takeda, and A. Benjebboury, "Transparent spectral confinement approach for 5G," in EuCNC 2017, Oulu, Finland, June 2017. [paper submitted].

[BKW+17b] J. Bazzi, K. Kusume, P. Weitkemper, Kazuki Takeda, and A. Benjebboury, "On resource grid for mixed numerologies in 5G," in in EuCNC 2017, Oulu, Finland, June 2017. [paper submitted]

[Bri98] E. O. Brigham. The fast Fourier transform, Prentice-Hall Inc, 1998.

[BTS+14] G. Berardinelli, F. M. L. Tavares, T. B. Sørensen, P. Mogensen, and K. Pajukoski, "On the potential of zero-tail DFT-spread-OFDM in 5G networks," in 80th IEEE Vehicular Technology Conference (VTC), 14-17 Sept. 2014.

[CCY+14] Aleksandra Checko, Henrik L. Christiansen, Ying Yan, Lara Scolari, Georgios Kardaras, Michael S. Berger and Lars Dittmann, "Cloud RAN for Mobile Networks - a Technology Overview," Communications Surveys & Tutorials, IEEE, vol. PP, no.99, pp.1,1, September 2014.

[DDF14] M. Danneberg, R. Datta, G. Fettweis, "Experimental testbed for dynamic spectrum access and sensing of 5G GFDM waveforms," IEEE Vehicular Technology Conference (VTC), 2014, pp. 1–5.

[DPG+16] D. D. Donno, J. Palacios, D. Giustiniano, and J. Widmer, “Hybrid analog-digital beam training for mmWave systems with low-resolution RF phase shifters,” in 2016 IEEE International Conference on Communications Workshops (ICC), May 2016, pp. 700–705.

[ETSI-NFV] European Telecommunications Standards Institute (ETSI): Industry Specification Group “Network Function Virtualization” (ISG NFV), http://www.etsi.org/technologies-clusters/technologies/nf

[FDP+17] M. C. Filippou, D. de Donno, C. Priale, J. Palacios, D. Giustiniano and J. Widmer, “Throughput vs. latency: QoS-centric resource allocation for multi-user millimeter wave systems,” to appear in proc. of the IEEE International Conference on Communications (ICC) , Paris, France, May 2017.

[FS98] H. G. Feichtinger, T. Strohmer. Gabor Analysis and Algorithms: Theory and Applications, Springer, 1998.

[GD09] R. H. Gohary and T. N. Davidson, “Non-coherent MIMO communication: Grassmannian constellations and efficient detection,” IEEE Trans. Inf. Theory, vol. 55, pp. 1176–1205, Mar. 2009.

[GVM+16] D. Garcia-Roger, J. F. de Valgas, J. F. Monserrat, N. Cardona, N. Incardona, “Hardware testbed for sidelink transmission of 5G waveforms without synchronization,” IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC), 2016.

[GZ16] J. Gebert, D. Zeller, „Fat pipes for user plane tunneling in 5G“, IEEE CSCN 2015, Berlin, October 2014

[HM00] B. M. Hochwald and T. L. Marzetta, “Unitary space-time modulation multiple-antenna communications in Rayleigh flat fading,” IEEE Trans. Inf. Theory, vol. 46, pp. 543–564, Mar. 2000.

[IETF-RFC2784] IETF RFC 2784, Generic Routing Encapsulation (GRE)

[IETF-RFC2890] IETF RFC 2890, Key and Sequence Number Extensions to GRE

[ITU13] ITU-Report, “Future spectrum requirements estimate for terrestrial IMT”, Dec. 2013.

[JS15] G. Jue, S. Shin. Keysight Technologies White Paper 2015: Implementing a flexible testbed for 5G waveform generation and analysis.

[JSV+14] S. Jha, K. Sivanesan, R. Vannithamby, A. T. Koc, “Dual Connectivity in LTE Small Cell Networks’’, IEEE Globecom Workshops, 8-12 December 2014, Austin, USA

[KBP+16] S R. Khosravirad, G .Berardinelli, K I. Pedersen, F .Frederiksen, “Enhanced HARQ Design for 5G Wide Area Technology”, VTC-Spring 2016

[KKD+15] F. Kaltenberger, R. Knopp, M. Danneberg, A. Festag, “Experimental analysis and simulative validation of dynamic spectrum access for coexistence of 4G and future 5G systems,” IEEE Vehicular Technology Conference (VTC), 2015, pp. 497–501.

[KYK+16] C. Kim, Y. H. Yun, K. Kim, and J. Y. Seol, “Introduction to QAM-FBMC: From waveform optimization to system design,” IEEE Communications Magazine 54 (11) (2016) 66–73.

[LJK+15] N. Lee, C. Jeong, J. Kim, and J. Park, “A new codebook structure for enhanced multi-user MIMO transmission in mmWave hybrid-beamforming system”, in 2015 IEEE Globecom Workshops (GC Wkshps), Dec. 2015.

[LPV+17] Y. Li, E. Pateromichelakis, N. Vucic, J. Luo, W. Xu, G. Caire, “Radio Resource Management Considerations for 5G Millimeter Wave Backhaul and Access Networks”, Communications Magazine, IEEE, June 2017 (to appear)

[MET216-D22] ICT-671680 METIS-II Deliverable D2.2, “Draft overall 5G RAN design,” June 2016.

[MET216-D31] ICT-671680 METIS-II Deliverable D3.1, “5G spectrum scenarios, requirements and technical aspects for bands above 6 GHz,” May 2016.

[MET216-D41] ICT-671680 METIS-II Deliverable D4.1, “Draft air interface harmonization and user plane design,” April 2016.

[MET216-D51] ICT-671680 METIS-II, Deliverable 5.1 Version 1, “Draft Synchronous Control Functions and Resource Abstraction Considerations”, May 2016.

[MET217-D52] METIS-II, “Final Considerations on Synchronous Control Functions and Agile Resource Management Framework for 5G“, March 2017.

[MET217-D62] ICT-671680 METIS-II, Deliverable 6.2 Version 1, “Final asynchronous control functions and overall control plane design”, April 2017.

[MH99] T. L. Marzetta and B. M. Hochwald, “Capacity of a mobile multipleantenna communication link in Rayleigh flat fading,” IEEE Trans. Inf. Theory, vol. 45, pp. 139–157, Jan. 1999.

[MMM16-W31] ICT-671650 mmMAGIC, white paper W3.1 Version 1.0, “Architectural aspects of mm-wave radio access integration with 5G ecosystem”, April 2016.

[NGMN15] NGMN Alliance Deliverable, “NGMN 5G White Paper”, Version 1.0, February 2015.

[NGMN16] NGMN Alliance Deliverable, “Description of Network Slicing Concept”, Version 1.0, January 2016.

[NNB+14] J. Nadal, C. A. Nour, A. Baghdadi, H. Lin, “Hardware prototyping of FBMC/OQAM baseband for 5G mobile communication systems,” IEEE International Symposium on Rapid System Prototyping, 2014, pp. 135–141.

[NNB16] J. Nadal, C. A. Nour, A. Baghdadi, “Low-complexity pipelined architecture for FBMC/OQAM transmitter,” IEEE Transactions on Circuits and Systems II: Express Briefs, 63 (1) (2016) 19–23.

[PDG+16] J. Palacios, D. De Donno, D. Giustiniano, and J. Widmer, “Speeding up mmWave beam training through low-complexity hybrid transceivers,” in 27th Annual IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC), Sept. 2016.

[PMD+15] E. Pateromichelakis, A. Maeder, A. De Domenico, R. Fritzsche, P. de Kerret and J. Bartelt, "Joint RAN/backhaul optimization in centralized 5G RAN," 2015 European Conference on Networks and Communications (EuCNC), Paris, 2015, pp. 386-390, June 2015.

[RBB+16] P. Rost, A. Banchs, I. Berberana, M. Breitbach, M. Doll, H. Droste, et al.: "Mobile Network Architecture Evolution toward 5G", IEEE Commun. Mag., May 2016

[RPW+16] C. Rosa, K. Pedersen, H. Wang et.al, "Dual Connectivity for LTE Small Cell Evolution: Functionality and Performance Aspects", IEEE Communications Magazine, vol. 51, no. 6, pp. 137-143, June 2016.

[Sam15] Samsung 5G vision white paper, Feb. 2015, available at: http://www.samsung.com/global/business-images/insights/2015/Samsung-5G-Vision-0.pdf

[SGA14] A. Sahin, I. Guvenc, H. Arslan, "A survey on multicarrier communications: Prototype filters, lattice structures, and implementation aspects," IEEE Communications Surveys Tutorials, 16 (3) (2014) 1312–1338.

[SHC10] T.-Y. Sung, H.-C. Hsin, Y.-P. Cheng, "Low-power and high-speed CORDIC-based split-radix FFT processor for OFDM systems," Digital Signal Processing, 20 (2) (2010) 511–527.

[SKD13] L. Szczecinski, S R. Khosravirad, P. Duhamel, "Rate allocation and adaptation for incremental redundancy truncated HARQ," IEEE Trans. Commun., vol. 61, no. 6,pp. 2580–2590, Jun. 2013.

[SSH+09] David Szczesny et al. "Performance analysis of LTE protocol processing on an ARM based mobile platform," 2009 International Symposium on System-on-Chip, Tampere, 2009, pp. 056-063.

[SSL02] P. Siohan, C. Siclet, N. Lacaille, "Analysis and design of OFDM/OQAM systems based on filterbank theory," IEEE Transactions on Signal Processing, 50 (5) (2002) 1170–1183.

[TGV+14] R. Trivisonno, R. Guerzoni, I. Vaishnavi, D. Soldani: "SDN-based 5G mobile networks: architecture, functions, procedures and backward compatibility", Transactions on Emerging Telecommunications Technologies (ETT), November 2014

[TPA11] Y. M. Tsang, A. S. Y. Poon, and S. Addepalli, "Coding the beams: Improving beamforming training in mmWave communication system," in IEEE Global Telecomm. Conf. (GLOBECOM 2011), Dec. 2011.

[VHM12] D. Van, W. Hendrickx, and T. Melia, "Method and apparatus for providing network access to a user entity," Dec. 12 2012, eP Patent App. EP20,110,290,264. [Online]. Available: http://www.google.com/patents/EP2533466A1?cl=en

[WBK+16a] P. Weitkemper, J. Bazzi, K. Kusume, A. Benjebboury and Y. Kishiyama, “Adaptive Filtered OFDM with Regular Resource Grid,” in 5G RAN design Workshop in conjunction with ICC 2016 IEEE, Kuala Lumpur, Malaysia, May 2016.

[WBK+16b] P. Weitkemper, J. Bazzi, K. Kusume, A. Benjebbour, and Y. Kishiyama, “On regular resource grid for filtered OFDM,” IEEE Communications Letters, 2016.

[WRC] http://www.itu.int/en/ITU-R/conferences/wrc/Pages/default.aspx

[YHZ+16] F. Yang, M. Huang, S. Zhang, S. Sirotkin.: “Radio Access Network Re-architecture to meet 5G Requirements – A SDN-based Paradigm Shift for Cellular Networks”, ICC 2016, 5G RAN Design WS, May 2016

[ZSW+15] Z. Zhao, M. Schellmann, Q. Wang, X. Gong, R. Boehnke, and W. Xu, “Pulse shaped OFDM for asynchronous uplink access,” in 2015 49th Asilomar Conference on Signals, Systems and Computers, 2015, pp. 3–7.

[ZT02] L. Zheng and D. N. C. Tse, “Communication on the Grassmann manifold: A geometric approach to the noncoherent multiple-antenna channel,” IEEE Trans. Inf. Theory, vol. 48, pp. 359–383, Feb. 2002.

A Further implementation aspects

A.1 Fast Fourier transform implementation

The IDFT operation may be computed through the IFFT. A typical implementation of the IFFT is the split-radix algorithm [SHC10], which yields identical computational cost for both the FFT and IFFT. The complexity can be computed by unfolding a complex multiplication into its real-valued counterparts (a complex multiplication is equal to four real-valued multiplications and two real-valued additions); then the total number of real operations is determined. In particular, the number of real-valued multiplications required by a FFT of size N is

$$C_{m_{FFT}}(N) = N\log_2 N - 3N + 4, \tag{A1}$$

the number of real-valued additions is

$$C\mathrm{a}_{\mathrm{FFT}}(N) = 3N\log_2 N - 3N + 4, \tag{A2}$$

And the total number of floating point operations (flops) equals

$$C\mathrm{f}_{\mathrm{FFT}}(N) = C\mathrm{m}_{\mathrm{FFT}}(N) + C\mathrm{a}_{\mathrm{FFT}}(N) = 4N\log_2 N - 6N + 8. \tag{A3}$$

Because of its importance for reducing the complexity costs of FBMC-OQAM, we propose the simultaneous calculation of the FFTs of size N of the two real-valued inputs involved, e.g. $g[n]$ and $h[n]$, via a single FFT of size N with complex-valued inputs, as shown in [Bri98]. After performing the IFFT, the additional total cost of reconstructing each individual FFT for both two real-valued inputs is

$$C\mathrm{m}_{\mathrm{C2R}}(N) = 8N, \qquad C\mathrm{a}_{\mathrm{C2R}}(N) = 4N, \qquad C\mathrm{f}_{\mathrm{C2R}}(N) = 12N. \tag{A4}$$

Note that from now on, $C\mathrm{m}$ refers to the number of real-valued multiplications, $C\mathrm{a}$ to the number of additions, and $C\mathrm{f}$ to the number of flops.

A.2 DFT spreading implementation

The DFT spreading operation is carried out for a number of samples *M*, generally much lower than *N*. It has the same cost as an FFT of size *M*.

A.3 Filter bank implementation through a polyphaser network

Multicarrier systems segment signals into (and reconstruct signals from) sub-bands via a set of N parallel filters called a filter bank. At the transmitter, the Analysis Filter Bank (AFB) is designed to split the signal into sub-bands according to the waveform properties. At the receiver, the Synthesis Filter Bank (SFB) is designed to rebuild the input signal by merging appropriately the outputs of these filters together.

We consider that these filter banks are designed as Finite Impulse Response (FIR) filters. It also follows from the concept of prototype filter that the filter banks are uniform, i.e. all filters in the filter bank are derived from the prototype filter via uniformly spaced frequency shifts. The uniformity of the filter bank is revealed by inspecting $P(z)$, the Z-transform of the FIR prototype filter $p[i]$ of KN coefficients, defined as

$$P(z) = \sum_{i=-\infty}^{\infty} p[i]z^{-i} = \sum_{i=0}^{KN-1} p[i]z^{-i} = \sum_{n=0}^{N-1} E_n(z^N)z^{-n}, \quad \text{(A5)}$$

where $z \in \mathbb{C}$, and $E_n(z^N) = \sum_{m=0}^{K-1} p[n+mN]z^{-mN}$ are the *polyphase components* of $P(z)$. Let us denote by $P_k(z)$ the k-th parallel filter of the filter bank, created by applying a frequency shift of k/N to $P(z)$

$$P_k(z) = P\left(ze^{-j2\pi k/N}\right) = \sum_{n=0}^{N-1} E_n(z^N)z^{-n}e^{-j2\pi k/N}. \quad \text{(A6)}$$

The most common implementation option for the design of the AFB and SFB is the PPN approach because it provides a significant reduction of the computational complexity through the use of polyphase filtering. A PPN filter bank modulates the prototype filter and carries out the addition, returning multiple bands of decimated and filtered time-domain data symbols from the IFFT stage. The IFFT implicitly multiplies by a complex phasor and is used to perform modulation to the different frequencies according to the prototype filter. The PPN is thus an efficient implementation of a filter bank because it removes the redundancy in the computations through time domain processing. Broadly speaking, this scheme requires an FFT of size N, roughly KN multiplications and storing KN samples for the PPN. As a disadvantage, a receiver implementing a polyphase network must perform subcarrier equalization in the time domain.

The PPN filtering structure is made clear by building $\boldsymbol{y}_s$, i.e. the N samples of the s-th symbol, as follows. Letting $\boldsymbol{p}_m$ be the vector whose n-th component is $p[n-mN]$, $\boldsymbol{p}_m$ is non-null only for $0 \le m \le K-1$, therefore

$$\boldsymbol{y}_s = \boldsymbol{x}_s \odot \boldsymbol{p}_0 + \boldsymbol{x}_{s-1} \odot \boldsymbol{p}_1 + \cdots + \boldsymbol{x}_{s-(K-1)} \odot \boldsymbol{p}_{K-1}, \quad \text{(A7)}$$

where $\boldsymbol{x}_m$ is the vector resulting from the IDFT of $\boldsymbol{X}_m$, $\boldsymbol{X}_m$ is the vector whose k-th component is X_{mk}, and $\odot$ denotes the Hadamard (element-wise) product. Note that, although creating the $\boldsymbol{p}$ vector requires phase rotations to get the linear phase filters at each subcarrier, the $\boldsymbol{p}$ vector is fixed and thus the computations need to be done just once. Producing $\boldsymbol{y}_s$ requires K point-wise complex products between $N \times 1$ vectors, and $K-1$ complex additions between $N \times 1$ pairs of vectors, so the total number of multiplications, additions and flops is

$$C\mathrm{m}_{\mathrm{PPN}}(N,K) = 4NK, \tag{A8}$$

$$C\mathrm{a}_{\mathrm{PPN}}(N,K) = 2NK + 2N(K-1), \tag{A9}$$

$$C\mathrm{f}_{\mathrm{PPN}}(N,K) = 2N(4K-1). \tag{A10}$$

If the prototype filter is symmetric in the frequency domain, as considered in this deliverable, the prototype filter samples are real, and hence the complexity in (A8)-(A10) can be reduced to

$$C\mathrm{m}_{\mathrm{PPN\text{-}\Re}}(N,K) = 2NK = \frac{1}{2} C\mathrm{m}_{\mathrm{PPN}}(N,K), \tag{A11}$$

$$C\mathrm{a}_{\mathrm{PPN\text{-}\Re}}(N,K) = 2N(K-1) = C\mathrm{a}_{\mathrm{PPN}}(N,K) - 2NK, \tag{A12}$$

$$C\mathrm{f}_{\mathrm{PPN\text{-}\Re}}(N,K) = 2NK + 2N(K-1) = \frac{2K-1}{4K-1} C\mathrm{f}_{\mathrm{PPN}}(N,K). \tag{A13}$$

B Single-waveform implementation complexity

In this annex, we will evaluate the complexity of the waveforms considered in Section 2.2.3. Assuming perfect synchronization between all the elements, at the transmitter only signal building procedures are accounted for, and at the receiver only the recovery of the signal is taken into account, and it is out of the scope of this deliverable the details of the operations needed to compute the equalizer coefficients, as well as the application of the equalizer itself. Thus, the computational complexity of the receivers and transmitters is the same. Note that for the CP-OFDM variants, the computation of the cyclic prefix actually has no computational cost because cyclic prefix insertion is equivalent to copying values already computed on memory.

B.1 Complexity of CP-OFDM

The CP-OFDM transmitter may be seen as an IFFT operation, and hence, the cost is

$$C\mathrm{m}_{\mathrm{TX/CP\text{-}OFDM}}(N) = C\mathrm{m}_{\mathrm{FFT}}(N) = N\log_2 N - 3N + 4, \quad \text{(A14)}$$

$$C\mathrm{a}_{\mathrm{TX/CP\text{-}OFDM}}(N) = C\mathrm{a}_{\mathrm{FFT}}(N) = 3N\log_2 N - 3N + 4, \quad \text{(A15)}$$

$$C\mathrm{f}_{\mathrm{TX/CP\text{-}OFDM}}(N) = C\mathrm{f}_{\mathrm{FFT}}(N) = 4N\log_2 N - 6N + 8. \quad \text{(A16)}$$

B.2 Complexity of W-OFDM

The W-OFDM transmitter may be seen as a CP-OFDM transmitter that introduces a stage after the cyclic prefix insertion where the multicarrier symbols are filtered by a short pulse shape ($K = 1$). Thus, its cost is

$$C\mathrm{m}_{\mathrm{TX/W\text{-}OFDM}}(N) = C\mathrm{m}_{\mathrm{FFT}}(N) + C\mathrm{m}_{\mathrm{PPN\text{-}}\mathfrak{R}}(N, 1) = N\log_2 N - N + 4, \quad \text{(A17)}$$

$$C\mathrm{a}_{\mathrm{TX/W\text{-}OFDM}}(N) = C\mathrm{a}_{\mathrm{FFT}}(N) + C\mathrm{a}_{\mathrm{PPN\text{-}}\mathfrak{R}}(N, 1) = 3N\log_2 N - 3N + 4, \quad \text{(A18)}$$

$$C\mathrm{f}_{\mathrm{TX/W\text{-}OFDM}}(N) = 4N\log_2 N - 4N + 8. \quad \text{(A19)}$$

B.3 Complexity of ZT-DFT-s-OFDM

The implementation of the ZT-DFT-s-OFDM would be the same as the one of the CP-OFDM waveform plus the cost of DFT spreading.

$$C\mathrm{m}_{\mathrm{TX/ZT\text{-}DFT\text{-}s}}(N,M) = N\log_2 N - 3N + M\log_2 M - 3M + 8 \quad \text{(A20)}$$

$$C\mathrm{a}_{\mathrm{TX/ZT\text{-}DFT\text{-}s}}(N,M) = 3N\log_2 N - 3N + 3M\log_2 M - 3M + 8 \quad \text{(A21)}$$

$$C\mathrm{f}_{\mathrm{TX/ZT\text{-}DFT\text{-}s}}(N,M) = 4N\log_2 N - 6N + 4M\log_2 M - 6M + 16. \quad \text{(A22)}$$

B.4 Complexity of P-OFDM

The P-OFDM transmitter may be conceived as a W-OFDM transmitter that allows the pulse shape to extend over the symbol period ($K \geq 1$). Modifying Eq. (A17)-(A19) accordingly, the cost is

$$C\mathrm{m}_{\mathrm{TX/P\text{-}OFDM}}(N,K) = C\mathrm{m}_{\mathrm{FFT}}(N) + C\mathrm{m}_{\mathrm{PPN\text{-}}\mathfrak{R}}(N,K) = N\log_2 N + 2NK - 3N + 4, \quad \text{(A23)}$$

$$C\mathrm{a}_{\mathrm{TX/P\text{-}OFDM}}(N,K) = C\mathrm{a}_{\mathrm{FFT}}(N) + C\mathrm{a}_{\mathrm{PPN\text{-}}\mathfrak{R}}(N,K) = 3N\log_2 N + 2NK - 5N + 4, \quad \text{(A24)}$$

$$C\mathrm{f}_{\mathrm{TX/P\text{-}OFDM}}(N,K) = 4N\log_2 N + 4KN - 8N + 8. \quad \text{(A25)}$$

B.5 Complexity of FBMC-QAM

The cost of FBMC-QAM is equivalent to that of P-OFDM, i.e.,

$$C\mathrm{m}_{\mathrm{TX/FBMC\text{-}QAM}}(N,K) = N\log_2 N + 2NK - 3N + 4, \quad \text{(A26)}$$

$$C\mathrm{a}_{\mathrm{TX/FBMC\text{-}QAM}}(N,K) = 3N\log_2 N + 2NK - 5N + 4, \quad \text{(A27)}$$

$$C\mathrm{f}_{\mathrm{TX/FBMC\text{-}QAM}}(N,K) = 4N\log_2 N + 4KN - 8N + 8. \quad \text{(A28)}$$

B.6 Complexity of FBMC-OQAM

As explained in Section A.1, the implementation that best fits the harmonization purposes of this deliverable consists of implementing the two real-valued IFFT of size N as one IFFT of size N, then adding the OQAM preprocessing, and finally performing two polyphase filtering procedures in parallel of size N.

The cost of computing one FFT of size N to compute two IFFTs of size N with the purely real and purely imaginary components as inputs, respectively, is

$$C\mathrm{m}_{\mathrm{FFT\text{-}OQAM}}(N) = C\mathrm{m}_{\mathrm{FFT}}(N) + C\mathrm{m}_{\mathrm{C2R}}(N) = N\log_2 N + 5N + 4, \quad \text{(A29)}$$

$$C\mathrm{a}_{\mathrm{FFT\text{-}OQAM}}(N) = C\mathrm{a}_{\mathrm{FFT}}(N) + C\mathrm{a}_{\mathrm{C2R}}(N) = 3N\log_2 N + N + 4, \quad \text{(A30)}$$

$$Cf_{\text{FFT-OQAM}}(N) = 4N\log_2 N + 6N + 8. \tag{A31}$$

The total cost of the FBMC-OQAM transmitter is

$$\begin{aligned} Cm_{\text{TX/FBMC-OQAM}}(N,K) &= Cm_{\text{FFT-OQAM}}(N) + 2Cm_{\text{PPN-}\mathfrak{R}}(N,K) \\ &= N\log_2 N + 4NK + 5N + 4, \end{aligned} \tag{A32}$$

$$\begin{aligned} Ca_{\text{TX/FBMC-OQAM}}(N,K) &= Ca_{\text{FFT-OQAM}}(N) + 2Ca_{\text{PPN-}\mathfrak{R}}(N,K) \\ &= 3N\log_2 N + 4NK - 3N + 4, \end{aligned} \tag{A33}$$

$$Cf_{\text{TX/FBMC-OQAM}}(N,K) = 4N\log_2 N + 8NK + 2N + 8. \tag{A34}$$

C Assumptions for Calculations of Functional Split Data Rates

Table C-6-1: Assumptions for Data Rate Calculations in Figure 4-3.

Parameter	Unit	5G AIV with LTE-like parameters		
Channel bandwidth	MHz	20		
OFDM subcarrier spacing	kHz	15		
FFT/IFFT size	Samples	2048		
Number of available subcarriers for transmission	-	1200		
Percentage of channel bandwidth used	%	90		
Number of transmit antenna ports	-	2	4	8
QAM modulation alphabet	-	64		
FEC code rate	-	0.75		
Number of spatial layers transmitted	-	2	4	8
Number of time-domain quantization bits in downlink direction	-	15		
Number of frequency-domain quantization bits in downlink direction	-	9		
Overhead introduced by the x-haul link. e.g. protection coding (>1)	-	1.33		

Table C-6-2: Assumptions for Data Rate Calculations in Figure 4-4.

Parameter	Unit	LTE	5G AIV < 6 GHz	5G AIV > 6 GHz
Number of antennas		32	64	64
Channel bandwidth	MHz	20	100	400
OFDM subcarrier spacing	kHz	15	30	120
FFT/IFFT size	Samples	2048	4096	4096
Number of available subcarriers for transmission	-	1200	3300	3300
Percentage of channel bandwidth used	%	90	99	99
Number of transmit antenna ports	-	8	32	64
QAM modulation alphabet	-	64	64	64

FEC code rate	-	0.75	0.75	0.75
Number of spatial layers transmitted	-	4	4	4
Number of time-domain quantization bits in downlink direction	-	15	15	15
Number of frequency-domain quantization bits in downlink direction	-	9	9	9
Overhead introduced by the x-haul link. e.g. protection coding (>1)	-	1.33	1.33	1.33

Additional Notes for Figure 4-3 and Figure 4-4:

- For LTE usual parameters were applied.
- For 5G AIV the channel bandwidth, the subcarrier spacing, and FFT size was varied, assuming to take fully account of channel bandwidth utilization (i.e. 99%) by applying filtering/windowing techniques (like UF-OFDM). In addition also the number of antenna ports was increased compared to LTE (32 and 64. resp.).
- The data rates were calculated for downlink. For uplink at split M3 higher data rates occur due to the transmission of soft-bits.