Call:H2020-ICT-2016-2

Project reference: 760809

Project Name:

E2E-aware Optimizations and advancements for Network Edge of 5G New Radio (ONE5G)

Deliverable D3.2
Recommended Multi-Service Performance Optimization Solutions for Improved E2E Performance

Date of delivery:	31/05/2019	Version:	1.0
Start date of project:	01/06/2017	Duration:	24 months

Document properties:

Document Number:	D3.2
Document Title:	Recommended Multi-Service Performance Optimization Solutions for Improved E2E Performance
Editor(s):	WINGS
Authors:	Daniela Laselva, Claudio Rosa, Jens Steiner, Klaus Pedersen (NOKIA), Nadège Varsier, Cesar Vargas Anamuro (ORANGE), Malo Manini, Nicolas Guérin (B-COM), Mustafa Emara, Miltiadis Filippou, Honglei Miao (INTEL), Elena Serna Santiago, Francisco Javier Lorca Hernando (TID), Isabel de la Bandera Cascales, Jesús Burgueño Romero, Jessica Mendoza Ruiz, Eduardo Baena Martínez, Sergio Fortes Rodríguez, David Palacios Campos, Raquel Barco (UMA), Gerhard Wunder, David Guzman (FUB), Galini Tsoukaneri (SEUK), Jimmy Jessen Nielsen, Nurul Huda Mahmood, Marco Centenaro, Beatriz Soret (AAU), Ioannis-Prodromos Belikaidis, Evangelos Kosmatos, Panagiotis Demestichas, Vera Stavroulaki, Andreas Georgakopoulos (WINGS), Mohamad Assaad, Koen de Turck, Ali Maatouk, Nesrine Ben Khalifa, Saad Kriouile (CNRS).
Contractual Date of Delivery:	31/05/2019
Dissemination level:	PU[1]
Status:	Final
Version:	1.0
File Name:	ONE5G_D3.2

Abstract

In this deliverable, we present our final recommendations for multi-service performance optimization solutions to achieve improved end-to-end (E2E) performance. Solutions included in this document are built on those presented in previous ONE5G report in D3.1. Use cases and Key Quality Indicator (KQI) framework described in D2.1 are also considered in this document. Solutions related to signaling and control plane optimizations, including mechanisms to optimize the User Equipment (UE) energy consumption and novel virtualization techniques are presented. Moreover, multiple radio resource configuration and allocation methods, and novel solutions for efficient usage of multi-channel access that helps improve the E2E performance are proposed. Finally, solutions related to spectrum aggregation, considering licensed and unlicensed bands, mobility optimizations, including elements of prediction, mobile edge computing assistance, and more fundamental mobility enhancements, and discovery methods, enhanced resource allocation techniques and device energy consumption aware methods in the context of device-to-device (D2D) communications are developed.

[1] CO = Confidential, only members of the consortium (including the Commission Services)

PU = Public

Keywords

5G new radio, channel estimation, compressive sensing, device-to-device, discontinuous reception, dual connectivity, end-to-end performance, licensed and unlicensed spectrum, load balancing, medium access control, multi-node connectivity, mobile edge computing, network architecture, optimization, predictive network control, pre-emption, radio resource control, radio resource management, resource allocation, scheduling, spectrum aggregation, user- and control-plane.

Executive Summary

In this public report, we present the final results of WP3 within ONE5G project. This WP is related to multi-service performance optimization solutions for improved E2E performance of the 3GPP New Radio. The material herein complements the material in D3.1 to represent the final findings from WP3. The developed solutions are largely applicable to both the Megacity and Underserved scenarios.

The first topic presented in this deliverable concerns the optimization of control plane signaling and power consumption of the user equipment (UE). For this purpose, a power consumption model of 5G New Radio (NR) is proposed, which extends the model presented in D3.1 [ONE18-D31]. Specifically, optimizations of the newly introduced RRC_INACTIVE radio resource control (RRC) state have been studied for machine-type communications (MTC) and it was found that the new RRC state can significantly improve network/service accessibility, service retainability and battery lifetime KQIs. Another study analyses the setting of DRX parameters when bandwidth parts (BWP) are used and finds that significant reductions in power consumption are achievable. Also, an innovative network-based device virtualization technique is proposed, where detailed analyses are performed regarding the optimal radio access network (RAN) split option for the scenarios of megacities and underserved areas. A compressed sensing based approach is proposed for reducing the amount of needed channel measurements in centralized radio access network (C-RAN).

The next topic addressed concerns various techniques for resource allocation to fulfill end-users' quality-of-service (QoS) requirements, including multi-service cases, and exploiting context awareness. These techniques are to a large extent complementary. Initially, various scheduling policies for improving the latency and hence service integrity KQI are proposed for distributed architectures, enabling a new use case on time sensitive networks (TSN). Also, efficient multiplexing of highly diverse services with significant QoS targets – such as enhanced mobile broadband (eMBB) and ultra-reliable low latency communication (URLLC) – is treated. A set of promising multi-cell scheduler enhancements for advanced mobile edge computing (MEC) / C-RAN architectures are developed to fully unleash the potential of using such architectures. Other investigated resource allocation techniques are multi-channel access (MCA) solutions for the 5G NR multi-service scenario. These techniques cover dual/multi-connectivity (DC/MC) for reliability and latency enhancement as well as carrier aggregation (CA) based on artificial intelligence (AI) for throughput enhancement.

Spectrum, connectivity and mobility optimizations are also studied. Regarding spectrum aggregation, an analysis of the available frequency bands that may be used for massive machine-type-communication (mMTC) in an unlicensed environment to achieve reliable coverage is performed. In a heterogeneous scenario, the use of a multi-cell aggregation scheduler (MAS) to optimize spectral efficiency and delay is recommended. In addition, an analysis of the coexistence of WiFi and licensed assisted access (LAA) in unlicensed bands is included. A DRS (Discovery Reference Signal) compensation method (CDRS) when disabling DRS signals is proposed. Finally, a study about the impact of listen-before-talk channel access on the reliability and latency performance in the 5 GHz unlicensed spectrum is performed and different solutions are analyzed. For the uplink, GUL (Grant-less Uplink) is recommended for low latency traffic over SUL (Scheduled Uplink).

Regarding mobility optimizations, context information about social events are recommended to be considered to improve performance. Also, the prediction of possible quality-of-experience (QoE) degradations could be used to anticipate the appropriate optimization actions in the network. Prediction techniques are also applied to improve URLLC. Mobility and access

management algorithm for optimizing different service classes across heterogeneous radio access technologies and specific scenarios such as high-speed trains are also presented, including optimization solutions for Vehicle-to-Everything (V2X) communication.

In the context of Device to Device (D2D) communications, studies from different perspectives are presented. The proposed solutions are mainly focused on radio resource allocation, energy consumption and challenges related to the exponentially increasing number of devices. Recommendations about binary power control, relay selection made at the user side, the use of the autonomous resource allocation mode and CC-HARQ (Chase Combining HARQ) are included.

Finally, Chapter 5 summarizes ONE5G WP3’s impact on 3GPP NR standardization specifying some candidate solutions to be included in future 3GPP NR releases such as 17 and 18.

Table of Contents

List of Figures

List of Tables

List of Acronyms and Abbreviations

ACK	Acknowledgement
AKA	Authentication & Key Agreement
ARQ	Automatic Repeat Request
AP	Access Point
BH	Backhaul
BS	Base Station
BW	Bandwidth
BWP	Bandwidth Part
CA	Carrier Aggregation
CAM	Cooperative Awareness Message
CC	Component Carrier
CCA	Clear Channel Assesment
CC-HARQ	Chase Combining HARQ
CDRS	Compensated Discovery Reference Signal
CG	Cell Group
CN	Core Network
CP	Control Plane
CoMP	Coordinated Multipoint
COT	Channel Occupancy Time
C-RAN	Centralized RAN
CQI	Channel Quality Indicator
CSI	Channel State Information
CSMA	Carrier Sense Multiple Access
CU	Centralized Unit
C-V2X	Cellular V2X
CW	Contention Window
D2D	Device-to-Device
DC	Dual Connectivity
DCI	Downlink Control Information
DRB	Data Radio Bearer
DRS	Discovery Reference Signal
DRX	Discontinuous Reception
DU	Distributed Unit
DVS	Device Virtualization Server
E2E	End-to-End

EAW	Event associated window
EIRP	Equivalent Isotropic Radiated Power
E-RAB	E-UTRAN Radio Access Bearer
ETSI	European Telecommunications Standards Institute
EP	End Point
FCC	Federal Communications Commission
FTAT	File Transfer Average Time
FTD	File Transfer Delay
FTP	File Transfer Protocol
GBR	Guaranteed Bit Rate
GRU	Gated Recurrent Unit
GUI	Graphical User Interface
gNB	next generation NodeB
GUL	Grant-less Uplink
HARQ	Hybrid Automatic Repeat Request
HetNet	Heterogeneous Network
HJB	Hamilton Jacobi Bellman
HLS	High Layer Split
HOF	Handover Failure
IoT	Internet of Things
ISD	Inter-Site Distance
ISDN	Integrated Services Digital Network
KPI	Key Performance Indicator
KQI	Key Quality Indicator
LAA	Licensed Assisted Access
LBT	Listen Before Talk
LLS	Low Layer Split
MAC	Medium Access Control
MAS	Multi-cell Aggregation Scheduler
MC	Multi-node Connectivity
MCA	Multi-Channel Access
MCOT	Maximum Channel Occupancy Time
MCS	Modulation and Coding Scheme
MDP	Markov Decision Process
MEC	Mobile Edge Computing
MeNB	Master LTE eNB
MgNB	Master 5G-NR gNB

MF	MulteFire
MFA	MulteFire Alliance
MIMO	Multiple Input Multiple Output
MOS	Mean Opinion Score
MTC	Machine Type Communications
mMTC	Massive MTC
MTD	Massive MTC Device
MU-MIMO	Multi-User MIMO
MU-PS	Multi-User Preemptive Scheduling
NACK	Non-Acknowledgement
NARX	Nonlinear Autoregressive exogenous models
NGC	5G New Generation Core
NR	New Radio
NR-U	NR-Unlicensed
OS	Operating System
PBCH	Physical Broadcast Channel
PDN	Packet Data Network
PF	Proportional Fair
PNC	Predictive Network Control
PDCP	Packet Data Convergence Protocol
PDU	Protocol Data Unit
PDCCH	Physical Downlink Control Channel
PDSCH	Physical Downlink Shared Channel
PSPACE	Polynomial SPACE
PPP	Poisson Point Process
PTS	Power Traffic Sharing
PUCCH	Physical Uplink Control Channel
PUSCH	Physical Uplink Shared Channel
PSM	Power Saving Mode
PSNR	Peak Signal Noise Ratio
PSTN	Public Switched Telephone Network
QoE	Quality of Experience
QoS	Quality of service
QSI	Queue State Information
RA	Random Access
RACH	Random Access Channel
RAI	Release Assistance Indicator

RAN	Radio Access Network
RB	Resource Block
RBP	Restless Bandit Problem
RLC	Radio Link Control
RLF	Radio Link Failure
RMSE	Root-mean-squared error
RNAU	RAN-based Notification Area Update
RNTI	Radio Network Temporary Identifier
ROMOS	Rate-based Objective Mean Opinion Score
RR	Round Robin
RRC	Radio Resource Control
RRH	Remote Radio Head
RRM	Radio Resource Management
RSRP	Reference Signal Received Power
RSRQ	Reference Signal Received Quality
RTS/CTS	Request-to-Send/Clear-to-Send
SIR	Signal to Interference Ratio
SFN	Single Frequency Network
SgNB	Secondary 5G-NR gNB
SoTA	State-of-The-Art
SON	Self-Organizing Network
SPS	Semi-Persistent Scheduling
SRB	Signaling Radio Bearer
SRS	Sounding Reference Signal
SS	Synchronization Signal
SSB	SS/PBCH Block
SUL	Scheduled Uplink
TAU	Tracking Area Update
TD-WPF	Time Domain Weighted Proportional Fair
TN	Transport Network
TSN	Time Sensitive Network
TTI	Time Transmission Interval
UE	User Equipment
UP	User Plane
UPF	User Plane Function
URLLC	Ultra-Reliable Low Latency Communications
V2X	Vehicle-to-Everything

VA	Venue-aware
VBR	Variable Bit Rate
VRU	Vulnerable Road User
WFO	Weighted Fair Opportunistic
WRC	World Radio Conference

1 Introduction

In this deliverable, we present our final recommendations for multi-service optimization solutions to achieve improved end-to-end (E2E) performance of the 3GPP New Radio (NR),,), where 3GPP release 15 is considered as the baseline. The current WP3 deliverable builds on the previous ONE5G report D3.1 [ONE18-D31], as well as on the use cases and Key Quality Indicator (KQI) framework as documented in D2.1 [ONE17-D21].

In line with the ONE5G project proposal, we present solutions for optimizing the E2E performance. The starting point is the QoS architecture and RAN protocol design with various options for new functionalities and different cross-layer optimizations as also outlined in Chapter 2 of ONE5G deliverable D3.1 [ONE18-D31]. For the sake of easy referencing, we here echo *what E2E means in the context of ONE5G*. The E2E paradigm presents the mobile network as a single pipeline between the application running in the UE/device and the remote host. Performance assessed in an E2E fashion is close to the experience of the end-user and reflects the behavior of the network and application as a whole. Besides the detailed RAN modeling, we target to capture the most relevant performance contributors being present in the E2E path outside the RAN in an abstract manner. Specifically, models of the impact with respect to latency and bandwidth (rate) originating in the fronthaul/backhaul network and the core are considered. E2E optimization is achieved either by designing enhancements and novel techniques or by configuring/tuning the values of the parameters relative to these techniques. We optimize the network towards improved E2E performance by optimizing directly or indirectly the Key Quality Indicators (KQIs) introduced in [ONE17-D21]. Better KQI scores imply improvements of the performance associated to the given KQI categories, and in turn, provisioning of enhanced end user experience in terms of objective QoE, which the KQIs are able to estimate. As explained in [ONE17-D21], the KQI framework distinguishes five main performance phases, each of them is associated to an individual KQI category: access to the network (*network availability* and *accessibility*), access to the service (*service accessibility*), and service quality or lack of it once the service is obtained (*service integrity* and *retainability*).

In principle, a KQI category could be built directly based on application-level performance indicators if those would be available. For example, rebuffering events would be key indicators of service integrity for video streaming services. However, in most cases assisting information from the UE/application client is not available at the network entities running optimization procedures. Alternatively, each KQI category can be built by aggregating relevant lower layer KPIs available at the network side, including the conventional *per-packet* QoS KPIs (latency, throughput, packet loss). Likewise, the final QoE score could be composed by integrating the KQI category scores weighted according to their relevance for a given service. A holistic hierarchy model is typically used to construct the required mapping from KPIs to KQI, and from KQIs to objective service QoE based on correlation rules as described in [LMK+18]. For the sake of E2E performance optimization in this deliverable, we will mainly aim at improving service-related KQIs (integrity, retainability, accessibility), assuming that the KQI scores rely on either direct application-level KPIs or lower layer (network/radio) KPIs. As further illustration, the two tables below show examples of service experience/quality through the KQIs (see also 3GPP Technical Report 32.862).

Table 1-1 Examples of KQIs for the eMBB service categories

KQI category	**File transfer**	**Video Streaming**	**Web-browsing**
Service accessibility (KQI-A)	Initial File Transfer Delay (s)	Video Streaming Start delay (s)	Page Response Delay (s)
Service Integrity (KQI-I)	File Transfer average throughput (Mbps) File Transfer delay (s)	Video Bit Rate (Mbps): Min/Average/Max (Assumes Adaptive Bit Rate)	Avg Page Throughput (Mbps) Page Download Time (s)

Service retainability (KQI-R)	File Transfer Cut off ratio	Video Streaming Stall: Frequency Number Time (s) Streaming cut off ratio	Page Transfer Cut-off Ratio

Table 1-2 Examples of KQIs for URLLC and eMTC service categories

KPI/KQI category	**KQI-sAccessability (KQI-sA)**	**KQI-sIntegrity (KQI-sI)**	**KQI-sRetainability (KQI-sR)**	**KQI-sMobility (KQI-sM)**
Basics	Initial message transfer delay	(E2E) Message Transfer Delay (E2E) Throughput (E2E) Reliability	Message Transfer Cut off	Mobility-related
Additional ones	Service range Service coverage	Positioning	Battery life	

The overall theme of Chapter 2 is signaling and control plane optimizations, including mechanisms to optimize the UE energy consumption. Firstly, we describe the adopted 5G NR power consumption model, which the work in this chapter is based on. Secondly, it is shown that especially the enhanced design of RRC protocol (with new RRC inactive state) and DRX framework offers promising opportunities for performance optimizations. Lastly, signaling and control plane optimizations in the form of novel virtualization techniques are presented, where especially network-based device virtualization techniques for C-RAN cases are developed, offering also opportunities for optimization of the physical layer control overhead from e.g. channel state information (CSI).

In Chapter 3, we present multiple radio resource configuration and allocation methods that help improve the E2E performance. Those include techniques that benefit from exploiting context-aware information. Promising schemes are presented for efficient scheduling of URLLC type of services, and for enhanced dynamic multiplexing of diverse services such as URLLC and eMBB on a shared channel, as well as solutions to fully unleash the potential of C-RAN network architectures (incl. multi-cell scheduling techniques). Secondly, Chapter 3 also presents novel solutions for efficient multi-channel access solutions; exploring enhanced carrier aggregation and multi-site connectivity schemes. In this context, solutions for achieving both enhanced throughput (say for eMBB type of services) and ultra-reliability (say for URLLC type of services) are developed, so essentially exploring both multi-cell data-split and data-duplication options. Among others, novel artificial intelligence (AI) solutions are presented for determining the best multi-channel access configuration for different devices depending on their service requirements, radio conditions, and network load conditions.

Chapter 4 is addressing spectrum, mobility, and D2D optimizations. Related to spectrum optimizations, the project has developed solutions for joint utilization of licensed and unlicensed bands, as well as standalone operation in unlicensed band(s). In particular, a KQI-based framework for mapping different user/services to certain spectrum chunks is presented, including utilization of the increased degrees of freedom that come with the higher bandwidth support by NR such as bandwidth part adaptation. In addition, considering a scenario with WiFi and LAA coexistence, a method called Compensation DRS (CDRS) is proposed to cover the lack of channel estimation when the used of Discovery Reference Signals (DRS) is not allowed, assigning the last stored MCS for that particular UE. This study is completed with an assessment of the different possibilities of accessing the non-licensed band channel in order to establish a framework for the optimization of different services based on its KQI. For unlicensed standalone operation, the performance of listen before talk (LBT) and GUL transmission in unlicensed spectrum is evaluated and different possible solutions are studied. Related to mobility, novel context aware QoE load balancing schemes are presented, including fundamental mobility enhancements (reduced handover interruption times, elements of UE autonomous decision makings, etc.). Context information about social events has been applied as an input for the optimization of the cellular networks. In addition, prediction methods can be used to improve network performance. In particular, prediction techniques have been applied to URLLC services in multi-connectivity scenarios and with the aim of anticipating possible QoE degradations and preventing optimization actions. Feature engineering techniques have been applied to improve prediction results. In particular, feature selection is used to select the most appropriate inputs for the prediction algorithms. Latency evaluations for V2X communications have also been carried out. Finally, handover performance in a high-speed train scenario has been studied. Regarding D2D optimizations, efficient D2D discovery methods, resource allocation techniques, and methods for minimize power consumption for mMTC devicesare developed. Cases with eMBB and mMTC types of traffic have been in the focus for the D2D related studies.

Each of the Chapters 2, 3, and 4 is concluded with a summary table, giving an overview of rich is the set of developed optimization solutions and associated recommendations, and how they help improve the E2E performance and related links to KQI metrics.

These summary tables are complemented by Chapter 5, where we specify how the work in WP3 has contributed to 3GPP standardization and which innovations can be considered for future 5G releases.

Finally, Chapter 6 concludes the report.

2 Control signaling and UE power consumption optimizations

The present chapter contains the contributions concerning control signaling and UE power consumption optimizations. Initially, the common NR UE power consumption model is presented in section 2.1, which is used in several works, both to optimize RRC and DRX parameters in section 2.2 and section 2.3, respectively, as illustrated in Figure 2-1. Thereafter, RRC optimizations for low layer control algorithms are considered. Finally, in section 0, device virtualization and RAN functional split aspects are considered. The device virtualization contribution studies how control signalling is impacted by moving UE functionality to the network edge and thereby enabling simplified UEs.

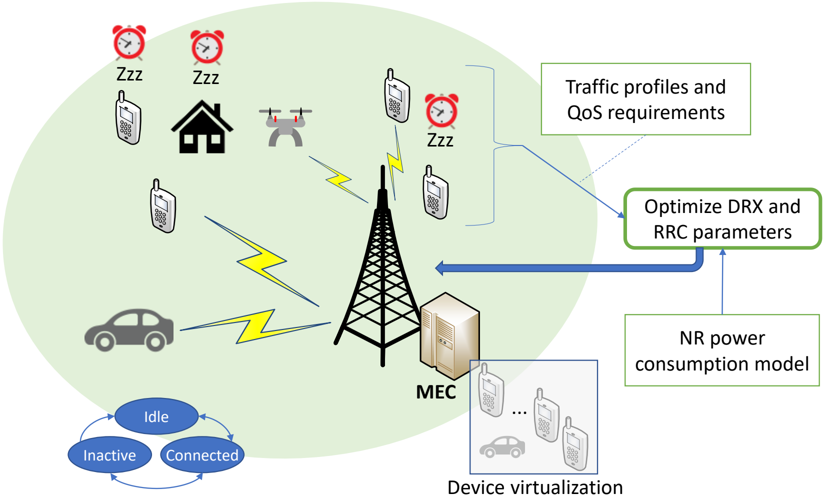

Zzz
Zzz
Zzz
Traffic profiles and QoS requirements
Optimize DRX and RRC parameters
NR power consumption model
MEC
Idle
Inactive
Connected
...
Device virtualization

Figure 2-1 Conceptual drawing of contributions in chapter 2 related to control signaling and UE power consumption optimizations

2.1 Adopted modeling for UE power consumption studies

2.1.1 UE power consumption model

In this section we describe the UE power consumption model adopted to carry out the UE power consumption studies presented in this deliverable. It extends the modelling proposed in D3.1 [ONE18-D31] with the aim to account more accurately for the effects that impact the consumption. For instance, we try to distinguish the consumption in relation to the type of transmission or reception, the transmit power, the number of active transceivers, and the used bandwidth. This is achieved by introducing new dedicated UE power states and power scaling. It is noted that this model has been accepted by the industry as part of the 3GPP Release-16 study on UE power savings in NR [3GPPTR-38840]. This model seems also rather in line with the findings from the latest measurements available based on LTE devices, e.g. [LAU14] [LBS+14].

Figure 2-2 illustrates the UE power consumption model with its UE power states and the corresponding state transitions. In Table 2-1, the revised parameters and parameter values of the model are given.

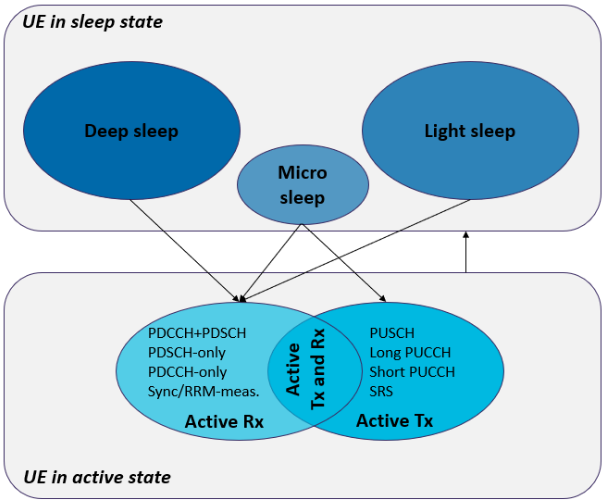


Figure 2-2 Illustration of the revised UE power consumption modelling with its states and corresponding transitions

Table 2-1 Extended UE power consumption model with updated parameter values, additional UE power states and power scaling as compared to D3.1 [ONE18-D31]. NB: all changes are marked with the italic font.

Power State		**Characteristics / Parameters**	**Relative Power** (per slot and assuming an arbitrary unit)
Sleep	**Deep Sleep**	The UE operates in its lowest power consumption mode, with baseband circuits maintaining timing to the lowest level of accuracy and minimal other baseband activities. RF circuits are not active.	***1*** *(Optional: 0.5)*
	Light Sleep	The UE has maintained timing using a clock and activity level which allows reception to be started with a reasonably small delay.	***20***
	Micro sleep	The UE enters this state if cross-slot scheduling is adopted, allowing to enter the sleep state immediately after monitoring PDCCH, i.e. without attempting decoding PDSCH.	***45***
Active RX	***PDCCH-only***	The RF receiver circuit is active, and the UE performs PDCCH decoding and micro-sleeps within the slot (i.e. no PDSCH and same-slot scheduling assumed). Label: P_{PDCCH}	***100***
	Synchronization / RRM measurements	The UE performs SSB or CSI-RS based RRM measurements (e.g. RSRP) and/or synchronization of the serving/camping cell.	***100***
	PDSCH-only slot	The UE is actively receiving or attempting to receive a signal on PDSCH (cross-slot scheduling). Valid for 4 Rx, 100 MHz.	***280***
	PDCCH + PDSCH	The UE is actively receiving or attempting to receive a signal on PDCCH and PDSCH.	***300***
	Antenna scaling	Scaling with the number of active Rx chains. The scaling is valid for 2 Rx. P_r is the power	$0.7 * P_r$

		consumption value when not scaled, depends on the UE power state, e.g. PDSCH-only.	
	BW scaling	Scaling with the active bandwidth. The scaling is valid for a bandwidth size W = 10, 20, 40, 80, 100 MHz. P_r is the power consumption when not scaled, depends on the UE power state, e.g. PDSCH-only.	$\left(0.4 + 0.6 * \frac{(W - 20)}{80}\right) * P_r$
Active TX	***PUCCH / PUSCH***	The UE is actively transmitting a signal. Long PUCCH or PUSCH transmissions are mainly assumed (rather than short PUCCH and SRS).	***250*** *(1 Tx chain)*
	Antenna scaling	Scaling with the number of active Tx chains (N_{Tx}). The scaling is valid to up to 2 Tx. *Pt* is the power consumption when not scaled, depends on the UE Tx power state, e.g. PUCCH/ PUSCH.	$1.4^{(N_{Tx}-1)} * P_t$
	Transmit power scaling	Scaling with the total transmit power (P_{Tx} in dBm). Power Amplifier efficiency (PA_{eff}) is assumed to be equal to 40%. *Pt* is the power consumption when not scaled, depends on the UE Tx power state, e.g. PUCCH / PUSCH.	$P_t + \frac{10^{\frac{P_{Tx}}{10}}}{PA_{eff}}$
	Short PUCCH/SRS	The UE is actively transmitting a short signal (short PUCCH and/or SRS). *Pt* is the power consumption when not scaled, depends on the UE Tx power state, e.g. PUCCH / PUSCH.	$0.6 * P_t$

Finally, any sleep state (deep, light, and micro sleep) is assumed to start with a ramp-down of the power used in the preceding active state and to conclude with a ramp-up of the power towards the following active state, as illustrated in Figure 2-3. For this purpose, additional power consumption and times associated with these transitions are given for each sleep state in Table 2-2.

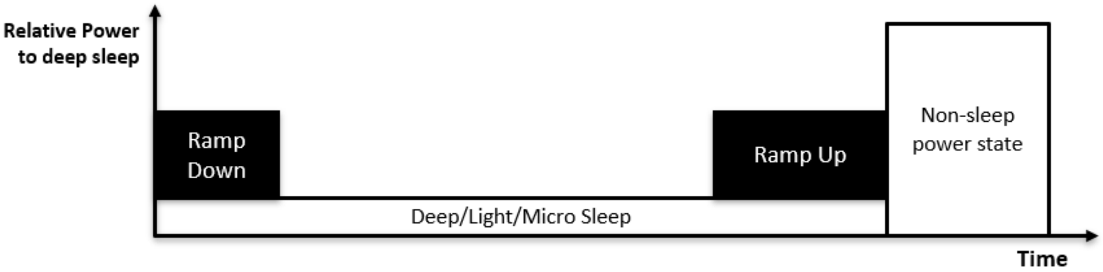


Figure 2-3 Illustration of the relative power in different UE power states including the additional ramping up and down transitions connected to the sleep states

Table 2-2 Additional transition power and transition time associated to the sleep states (when transiting from and to non-sleep states)

Sleep state	**Additional UE Power Consumption for State Transition (*)** (for ramp-up + ramp-down)	**Total transition time** (for ramp-up + ramp-down)
Deep Sleep	450	20 ms

Light Sleep	120	6 ms
Micro sleep	0	0 ms (**)

(*) Relative power per 1 ms.

(**) Immediate transition assumed.

We remark that we have contributed to shape this UE power model in 3GPP (see e.g. [3GPPR1-1811478]) and we are disseminating it in a conference paper for IEEE VTC Fall, see [MLF+19], where further details of the model are discussed.

2.1.2 DRX operations in NR

In this section we recollect the DRX operations defined in NR, which are used in the studies presented in this delievrable for additional UE power saving. DRX is set by configuring the timers and parameters given in Table 2-3. It is noted that the term NR unit below refers to the scheduling time unit [ONE18-D31].

Table 2-3 DRX parameters in NR

DRX Parameter in NR	*Description*	*Label*
drx-InactivityTimer	The number of consecutive NR unit(s) after the scheduling slot in which a PDCCH indicates an initial UL or DL user data transmission for the MAC entity. Unit in milliseconds.	T_i
drx-onDurationTimer	The number of consecutive NR unit(s) at the beginning of a DRX Cycle. Unit in milliseconds. For RRC_Idle, this corresponds to the paging monitoring occasions within a DRX cycle.	T_{on_c}
		T_{on_i}
drx-Cycle	DRX cycle for RRC_Idle UEs.	T_{pi}
drx-ShortCycle	Short DRX cycle. Unit in milliseconds.	T_{sc}
drx-ShortCycleTimer	The number of consecutive NR unit(s) the UE shall follow the Short DRX cycle.	T_{sc_t}
drx-LongCycle	Long DRX cycle. Unit in milliseconds.	T_{lc}
drx-RetransmissionTimerDL (UL)	Per DL HARQ process: the maximum number of consecutive NR unit(s) until a DL (UL) retransmission (grant) is received.	N_r
RRC Release/Suspend timer	The number of consecutive NR unit(s) after the scheduling slot in which a PDCCH indicates an initial UL or DL user data transmission for the MAC entity, after which the network determines UE should move out of RRC_Connected to RRC_Idle mode (via an RRC connection release procedure) or to RRC_Inactive (via an RRC connection suspend procedure). Unit in milliseconds.	T_r

It is noted that in [MLF+19], we have evaluated different DRX configurations for NR, including the use of short and long DRX, to identify the best Rel-15 baseline. Results have shown that proper configuration of DRX including short DRX can reduce the UE power consumption by 30% and the average latency by 70%, as compared to not using short DRX. In [MLF+19], we have also investigated further UE power saving potential by using other Rel-15 features such as cross-slot scheduling and PDCCH blind decoding with reduced candidate set, as well as the Rel-16 enhancements of go-to-sleep and wake-up signalling in the scope of the related 3GPP initiative of UE power saving in NR [3GPP38.840].

2.1.3 Application of the UE power consumption model

The UE power consumption model given in Section 2.1.1 is needed along with the DRX modeling provided in Section 2.1.2 to analyze the power consumption impact of different Radio resource Management (RRM) strategies including DRX and RRC state handling configurations. In the following, we describe how they can be used to compute the UE power consumption when the UE is in different RRC states.

Firstly, we consider RRC_IDLE and RRC_CONNECTED states. When the UE is in RRC_IDLE, the energy consumption during a DRX cycle when no paging message is received, is given as:

$$E_{idle} = \left(T_{pi} - T_{on\_i}\right) * P_{ds} + T_{on\_i} * P_{PDCCH}$$

where the timing parameters are defined in Table 2-3 DRX parameters in NR, power parameters are defined in the power model in Section 2.1.1, P_{ds} is the relative power consumption in the deep sleep state, and P_{PDCCH} is the power consumption to check PDCCH. If the device is in the RRC_IDLE state and wants to transmit data, it needs to move to the RRC_CONNECTED state. This state switch includes the random access process and part of the attach process (at least the Packet Data Network (PDN) connectivity setup). The energy consumption for the state switch is given as:

$$E_{switch} = T_{RA} * P_{RA} + T_{AT} * P_{AT}$$

where T_{RA} is the time required to complete the Random Access process, P_{RA} is the average power usage for the random access process, T_{AT} is the time required to complete the attach process (excluding Authentication & Key Agreement (AKA)), and P_{RA} is the average power usage for the attach process.

In RRC_CONNECTED state, data reception starts once PDCCH indicates a new transmission from the gNB, and the UE starts the DRX inactivity timer. The energy consumption is defined as:

$$E_{data} = T_{data} * P_{at,CA-N} + (T_i - T_{data}) * P_{PDCCH}$$

where $P_{at,CA-N}$ is the power consumption in the active state to transmit data, and can be replaced with $P_{ar,\,CA\text{-}M}$ (power consumption in the active state to receive data) if it is data reception (ONE5G D3.1 [ONE18-D31]). In the RRC protocol of Rel. 15, the device remains in the RRC_CONNECTED state for a period of time defined by the RRC release timer before switching to the RRC_IDLE state and starting the short DRX cycle timer. The energy consumption is defined as:

$$E_i = T_i * P_i$$

where T_i is the time spent in the RRC_CONNECTED state after a successful data transmission or reception, and P_i is the average power consumption during that period of time. Once the RRC release timer expires, the UE moves to short DRX cycle and short DRX cycle timer starts. The energy consumption is defined as:

$$E_s = T_{sc\_t} * \frac{\left(\left(T_{sc} - T_{on_c}\right) * P_{ls} + T_{on_c} * P_{PDCCH}\right)}{T_{sc}}$$

After the short DRX cycle timer expires, the UE moves to long DRX cycle and the RRC inactivity timer T_{tail} starts. The energy consumption is:

$$E_{tail} = T_r * \frac{\left(T_{lc} - T_{on\_c}\right) * P_{ls} + T_{on\_c} * P_{PDCCH}}{T_{lc}}$$

Therefore, the overall energy consumption for one RRC establishment/release is given as:

$$E_{total} = N_{idle} * E_{idle} + E_{switch} + E_{data} + E_i + E_s + E_{tail}$$

where N_{Idle} is the number of DRX cycles in RRC_IDLE/RRC_INACTIVE before transition to RRC_CONNECTED.

The above analysis outlines that a large percentage of the total energy consumption depends on the state that the device is in, the time it spends on that state, as well as the frequency with which it switched between different states. Therefore, the state that the device switches to must be carefully selected to result in the lowest energy consumption possible. For example, devices such as smartphones might require frequent connections to the network, and as such, the power consumption for performing the complete connection establishment process from the RRC_Idle state can incur a significant energy cost. In these cases, keeping the device partially connected to the network with the RRC_Inactive state can have a positive impact in their total energy consumption. On the other hand, for devices such as IoT that exchange data infrequently, the power cost of performing the complete connection establishment process may not be significant as this procedure is followed only a few times a day. As such the use of the RRC_Idle state is preferred, as not only it does not increase the power usage, but it also saves network resources.

The device mobility is another important characteristic that need to be considered for the device power consumption. Although the RRC_Inactive state can decrease the power consumption for establishing a connection, this is mainly true when the connection is resumed to the same base station. However, if the connection is resumed to a new base station, information regarding the previously established connection must be transferred from the old to the new base station. This not only increases the network access delay, but also requires the repetition of the complete connection establishment process (similarly to being in the RRC_Idle state) if the connection resumption is rejected, thus significantly increasing the power consumption. As such, for highly mobile devices that connect to a new base station for each data transmission, the use of the RRC_Idle state can present benefits in terms of total energy consumption and network access delay.

Another characteristic that needs to be considered is whether devices transmit their data in one go. As mentioned earlier, after a device transmits its data, it starts the RRC release timer, during which it remains in the RRC_Connected state. If no further data needs to be exchanged before the expiration of the RRC release timer, the device is switched to the RRC_Idle/RRC_Inactive state. The duration of the RRC release timer is network operator specific (10-50 seconds in most commercial networks). For devices that have intermittent data transmissions (e.g. browsing) the use of the RRC release timer can significantly reduce their energy consumption, as it avoids the need to establish a new connection for each data transmission. However, for devices that exchange all of their data at once (e.g. IoT sensors/meters), remaining in the RRC_Connected state after the data transmission wastes energy unnecessarily. Although the DRX functionality can be used in the RRC_Connected state, only short cycles can be applied, meaning that the device still needs to frequently check the PDCCH for network notifications. Therefore, the traffic patterns of the device need to be taken into account when assessing the energy consumption, and allow the device to unilaterally release its connection (e.g. with the use of RAI [3GPP24.301]), or use short RRC release timers to quickly switch to low power states.

Finally, the probability of receiving data from the network while in the RRC_Idle or RRC_Inactive states should also be considered to decide the DRX cycle values, which in turn affect the energy consumption. For example, devices such as smartphones are highly likely to receive network originated data at any time, while autonomously operated devices such as IoT, might be receiving downlink data scarcely. Based on these characteristics, short DRX cycles must be used for some devices, while long DRX cycles can be beneficial for other device to prevent them from unnecessarily waking up to check the PDCCH.

To conclude, our energy consumption model provides insights of how the energy is consumed during a complete connection cycle of a device (i.e. connection establishment, data exchange, connection release, light/deep sleep), and helps to identify the procedures with the highest energy cost. As devices’ traffic and operation characteristics vary significantly, it is recommended that future energy optimization approaches are based on them to a great extent.

2.2 Service-dependent optimized RRC state handling in NR

The RRC state of a UE plays a crucial role in defining the service and end-user performance. For instance, it affects the UE reachability and power consumption by determining the operations the UE needs to perform during the current RRC state and during potential RRC state transitions (comprising control-plane and mobility procedures). In this section, we discuss how the RRC state transition handling could be optimized to exploit the new RRC_Inactive state, introduced as part of the NR RRC state machine in Release-15. It is noted that we are disseminating these optimizations in a conference paper for IEEE PIMRC 2019, see [KLR19].

The RRC state transitions are determined by the network by means of network timers, denoted as RRC release/suspend timer, which control the connection release to RRC_Idle or suspend to RRC_Inactive. However, it is challenging for the network to determine the optimal RRC state transition, and hence to set these timers appropriately. Among other factors that influence the timers' configuration, the traffic type and traffic profile (or data activity) of the UE are key. For instance, the RRC_Inactive design is particularly tailored to traffic types, such as eMTC, that present regular or sporadic data exchanges, which may be relatively widely scattered in time. Similarly, a large number of packets in eMBB services have also a small payload and may be infrequent, such as background and keepalive messages, and could benefit from the RRC_Inactive state as well. However, depending on the suspend timer setting, the sporadic activity may result either in seldom directing a UE out of the RRC Connected state (if the timer is set too long as compared to the data interarrival timing) or in causing frequent and subsequent RRC resume/suspend ping-pongs (if the timer is set too short vs. the data interarrival timing), which are undesired. Frequent RRC transitions should be avoided because of the additional signalling and battery consumption and the need to perform the random-access procedure to resume the connection. If a large number of MTC devices attempt resuming at the same time, the likelihood of collision on the RACH channel will increase too, leading to increased control plane latency. On the other hand, when the UE is in RRC Connected state, one major cost for the network is the mobility related signalling to move the UE to a target cell through the measurement-assisted handover procedures. The signalling is costly since it involves signalling over the Uu, X2/Xn, and 5G New Generation Core (NGC) interfaces, and therefore should be limited.

In D3.1, an optimized RRC state transition framework was proposed to exploit the new flexibility of NR, with the aim of controlling the trade-off between UE power consumption, latency and signalling overhead, while complying efficiently with services with diverse characteristics [ONE18-D31]. Based on the underlying considerations above, the framework comprises different strategies to determine the need for an RRC state transition as a function of the following factors:

1) *UE data activity*: When detecting data inactivity during a period (controlled by the RRC suspend-timer), the network may move a UE from RRC Connected to the RRC_Inactive state by sending an RRC suspend message. In case of frequent data arrival, the network keeps the UE in RRC connected mode, achieving UE power savings by means of DRX.

2) *Network load*: Determining the number of UEs to be kept in RRC Connected mode has to account for the fact that control plane resources used by RRC Connected UEs are limited. Specifically, the number of simultaneous RRC connections within a cell and the usage of system radio resources such as PUCCH / RACH preambles should not become a bottleneck.

3) *User mobility*: To avoid the signalling cost associated to mobility, the UEs can be moved to RRC_Inactive or RRC_Idle state, where the mobility is handled locally at the UE through cell reselection procedures, with the exception of the procedures of RAN-based Notification Area Update (RNAU) and/or Tracking Area Update (TAU). To this end, the frequency of RNAU procedures could be the major differentiator to determine whether the RRC_Inactive or RRC_Idle state should be used.

In order to optimally operate according to the high-level strategies introduced above, it remains to quantify the following aspects related to traffic and mobility:

- *From the traffic point of view*: Under which payload size and payload arrival frequency is it useful to consider RRC_Inactive rather than RRC_Connected or RRC_Idle?
- *From the mobility point of view*: Under which mobility profile is it useful to consider RRC_Inactive rather than RRC_Connected or RRC_Idle?

In the following, we describe the steps and associated methodology adopted to perform such evaluation, towards quantifying the key advantages of using RRC_Inactive and the proposed RRC state transition framework.

First, the evaluation is performed for the single UE scenario, with the aim to understand the fundamental trade-off between UE power consumption, control-plane (CP) and user-plane (UP) latency, and network signalling. Therefore, the KQIs in focus are, on one hand, network and service accessibility (KQI-nA, KQI-sA) covering, for instance, the control-plane latency for initial access (i.e. time to access the network/server before being able to transmit/receive the first UP packet). On the other hand, service retainability is in focus as well covering the UE battery life (KQI-sR). Finally, service integrity (KQI-sI) is considered, covering the user-plane latency (i.e. time to transmit/receive a packet). For further details on the KQI framework, the reader can refer to our dissemination paper [LMK+18].

Table 2-4 compares various RRC state transitions under stationary scenarios, comprising the cases where the UE needs to perform the first uplink (UL) data transmission when is currently in RRC_Idle (baseline #1), RRC Connected (baseline #2), or RRC_Inactive. Various assumptions are considered related to the UE processing delay (to be equal or less than the values assumed in [3GPP36.912]) and the Transmission Time Interval (TTI) length. It is observed that the CP latency prior to transmitting the first UL payload is significantly reduced for UEs in the RRC_Inactive state as compared to RRC Idle. This is achieved thanks to three main contributions: i) reduced number of control-plane messages required prior to the transmission, ii) reduced processing delay at the UE to process the RRC messages, and iii) reduced TTI size, leading to shorter transmission time. It is noticed as well, that obviously RRC_Inactive does not bring advantages as compared to RRC_Connected, as no mobility signalling (and associated UE power) nor other typical UE reporting are assumed in this (stationary) scenario. The RRC_Inactive state achieves ~60% and ~86% reduction in control-plane latency at RRC state transition to RRC_Connected, as compared to RRC_Idle, for case b and c respectively.

Table 2-4 RRC state comparison in terms of CP latency, UE power consumption and network signalling prior to the first UL data transmission (i.e. including the scheduling grant for the transmission) for a UE in RRC_Idle, RRC_Inactive and RRC_Connected state.

Case	RRC state transition prior to the first UL data transmission	Normalized Power consumption (-)	Minimum state transition CP latency (ms)	Total # NW signalling messages (RAN+CN) (-)
a	**From NR RRC_Idle to RRC_Connected** (baseline #1)	1.0	76.00	12 (9+3) [Fan5GD4.2]
b	**From NR RRC_Inactive to RRC_Connected** (Rel-15; UE processing delay according to [3GPP36.912]; TTI = 1 ms)	0.43	31.50	5 (5+0)
c	**From NR RRC_Inactive to RRC_Connected** (Rel-15; UE processing delay reduced to a fifth vs. [3GPP36.912]; TTI = 1 ms)	0.22	13.00	5 (5+0)
d	**From NR RRC_Inactive to RRC_Connected** (Rel-15; UE processing delay reduced to a fifth vs. [3GPP36.912]; short TTI = 0.143 ms)	0.05	8.07	5 (5+0)

e	**No state transition: UE in NR RRC_Inactive** (Rel-16+: UL data transmission along with so-called message #3 (MSG3); reduced processing delay, short TTI as in d)	0.05	1.86	2 (2+0)
f	**No state transition: UE in NR RRC_Connected** (Rel-15; UL data transmission after Scheduling Request and scheduling grant; reduced UE processing delay and short TTI as in d). Baseline #2.	0.05	1.86	2 (2+0)

It is noted, that the UE power consumption and CP latency given in Table 2-4 are computed by assigning the corresponding power consumption tag and latency tag to each operation required prior to the payload transmission (e.g. UE message transmission, processing, or reception). The power tags are taken from the UE power consumption model given in Section 2.1, and depend on the UE power state.

Table 2-5 shows the CP latency components of an RRC state transition from RRC_Inactive to RRC_Connected, comprising cases c and d in Table 2-4, for which the total CP latency is around 10 ms. Table 2-5 highlights the impact of the TTI size, showing that with the mini-slot of 2-ODFM symbols (i.e. TTI size = 0.143 ms; 15 kHz subcarrier spacing, SCS, is assumed), the total CP latency can be reduced to 8 ms (as compared to 13 ms when using the full slot of 1 ms).

Table 2-5 CP latency component split for the RRC state transition from RRC_Inactive to RRC_Connected when varying the TTI length

Component	Description	14 OFDM symbols (TTI = 1ms)	7 OFDM symbols (TTI = 0.5 ms)	2 OFDM symbols (TTI = 0.143)
1	Delay due to RACH scheduling period (1 TTI period), (i.e. avg. delay to nearest TTI)	0.5 ms	0.25 ms	0.0715 ms
2	**UE's transmission** of **RACH Preamble**	1 ms	0.5 ms	0.143 ms
3	Preamble detection and processing in gNB + Transmission of Random Access Response incl. **UE's reception of scheduling grant** and timing adjustment	1.5 ms	0.75 ms	0.643 ms
4	**UE Processing Delay** (decoding of scheduling grant, timing alignment and C-RNTI assignment + L1 encoding of RRC Connection Resume Request)	1 ms	1 ms	1 ms
5	**UE's transmission of RRC Connection Resume Request (MSG3)**	1 ms	0.5 ms	0.143 ms
6	**gNB processing delay** (RRC and L2)	3 ms	3 ms	3 ms
7	**gNB's transmission of RRC Connection Resume** (and UL grant)	1 ms	0.5 ms	0.143 ms
8	**UE processing delay** (RRC and L2)	3 ms	3 ms	3 ms
9	**UE's transmission of RRC Connection Resume Complete** (MSG5, including NAS Service Request)	1 ms	0.5 ms	0.143 ms
Total CP latency		13 ms	10 ms	8 ms

The second part of the evaluation is presented in the following. The optimization of the RRC state transition is investigated with system level simulations when employing DRX as an additional key mechanism for UE power saving. The investigations are made in a dense urban scenario and under the DRX configurations defined in [3GPPTR-38840], focusing on the impact of the traffic profile. It is noted that DRX is employed in each RRC state, comprising connected-mode DRX (C-DRX), which determines the PDCCH monitoring, as well as inactive/idle-mode DRX (I-DRX), which determines the paging monitoring.

In this study, we focus on downlink and assume as baseline that UEs are kept in RRC_Connected mode with a DRX cycle of 160 ms, with no RRC state transition allowed. The performance of RRC_Inactive under different configurations in terms of paging cycles is investigated for a fixed RRC connection suspend timer of 10 ms. For further details and scenarios, the reader can refer to our conference paper for IEEE PIMRC 2019, see [KLR19].

Figure 2-4 shows the average reduction of UE power consumption achieved by RRC_Inactive relatively to the RRC_Connected-only baseline, as a function of the user packet rate. Similarly, Figure 2-5 shows the average percentage increase of the total packet delay as compared to the RRC_Connected-only baseline. The total packet delay may comprise both the control plane delay component (if an RRC state transition occurred), and the user plane delay for the actual scheduling and transmission. The additional case, when the RRC state transition to RRC_Idle is allowed after the expiration of the RRC release timer, is considered too. The RRC release timer for this case is set to 1s, in line with the typical range used in LTE (i.e. 1-10 s).

It is observed from the figures how RRC_Inactive can achieve a significant power reduction at a slight cost of latency increase when the packet rate is low (e.g. 2 packets/s or below). At higher packet rates, instead, RRC_Inactive is less desirable because the achieved power consumption reduction is not significative, whereas the resulting increase in latency becomes large. For such cases, RRC_Connected should be preferred. We remark that the latency is mainly affected by the paging cycle length at low packet rate levels because a UE is likely to be moved back to RRC_Inactive state after each packet transmission was made in RRC_Connected. Instead, at higher packet rates, it is more likely that a UE remains in RRC_Connected after a transmission is made since a new packet may appear in the buffer. This assumes, though, that the RRC connection suspend timer is sufficiently large, which is not the case in these simulations (fixed value of 10 ms is assumed). This effect, however, was confirmed by observing the number of RRC state transitions occurring at different packets rates while varying the timer value, see [KLR19].

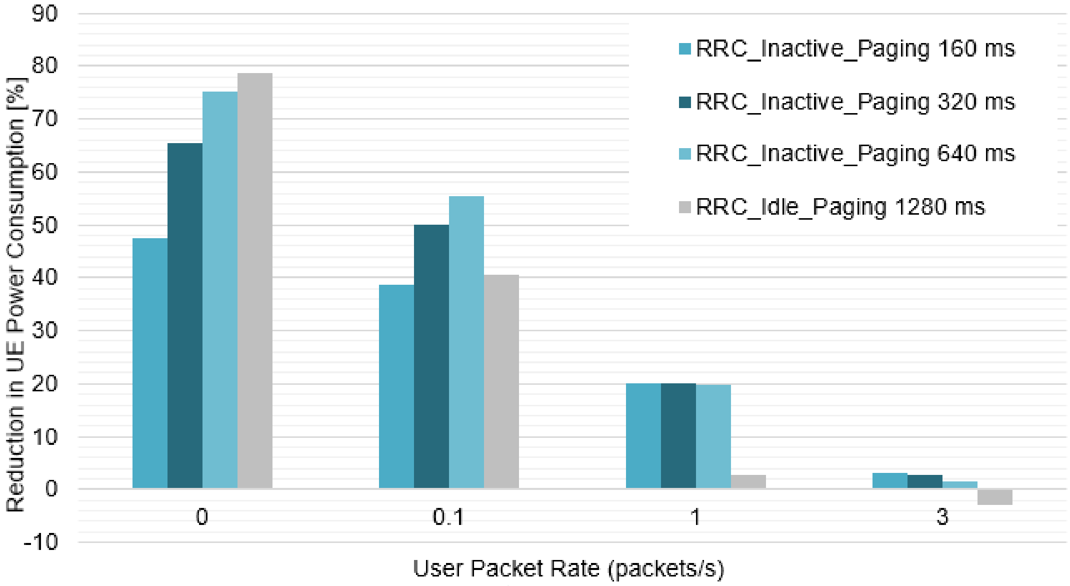


Figure 2-4 Preliminary example of the impact of the inter-arrival packet time to the CP latency prior to the first transmit UL payload

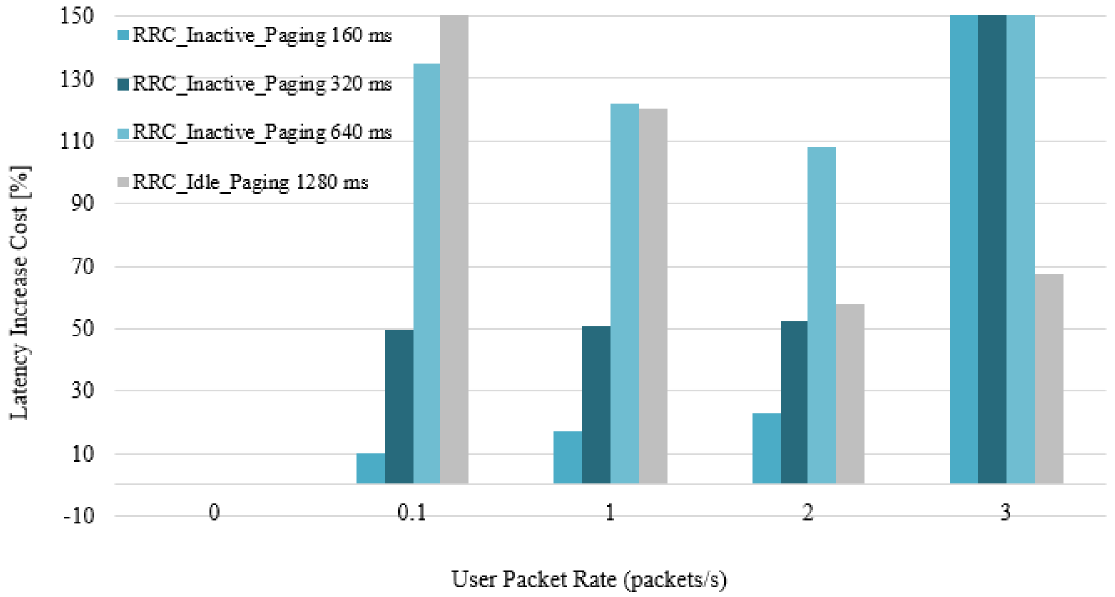


Figure 2-5 Latency increase cost vs. the RRC-Connected-only baseline as function of the user packet rate. NB: the KPI is upper bounded to 150% for readability.

Based on the results given above, it is apparent that optimizing the RRC state handling exploiting the RRC_Inactive state is rather beneficial. Numerically, RRC_Inactive can lead to reduced network/service accessibility, achieving up to 89% shorter control-plane latency at RRC state transition to RRC_Connected, as compared to RRC_Idle, i.e. 8 ms vs 76 ms. Further, RRC_Inactive can lead to higher service retainability, achieving ~70% longer battery life compared to RRC_Connected in no data scenarios, and ~40% extended battery life for infrequent packet arrival. In addition, it can achieve good service integrity, with an overall low latency, where the latency increase compared to RRC_Connected can be limited to ~10% for infrequent traffic.

Therefore, we recommend optimizing the RRC state handling as follows. First, the RRC state handling should be optimized as a function of the service/traffic requirements (e.g. packet inter-arrival rate at a UE and QoS requirements), by setting properly the RRC connection suspend timers. Longer timer settings (hundreds of millisecond or seconds) can be used when more frequent traffic is generated, to retain the UE in RRC_Connected state and avoid frequent RRC state transitions, which cause latency and may lead to RACH collisions. Vice-versa, shorter timer settings (down to few tens of millisecond) can be used for infrequent traffic, allowing to timely move a UE to RRC_Inactive benefiting from its power saving properties. The paging cycle for UEs in RRC_Inactive can be set to small values (in order of few hundreds of milliseconds) to keep the latency penalty due to DRX sleep periods low, since no significant advantage in UE power is achieved at larger cycles. It is noted that in future 3GPP releases, the RRC Inactive state is expected to support small data transmission, i.e. the transmission of small payload in RRC_Inactive without requiring first an RRC state transition to RRC_Connected. This will extend the applicability of RRC_Inactive also at higher traffic rates as long as the payloads to transfer are small enough to be accommodated without requiring a transition to RRC_Connected (e.g. up to 1000 bits).

Furthermore, the optimization of the RRC state handling should consider the mobility profile of a UE as well, attempting to limit the mobility related signalling to the RAN / CN which may be required for UEs in RRC_Connected. Specifically, RRC_Connected with long DRX can be used for (semi-)stationary UEs having medium/high traffic frequency. Limited mobility related signalling and UE power savings are achieved thanks to DRX and because of the UE stationarity state.

Instead, RRC_Inactive with longer DRX can be used for infrequent traffic and low/medium mobility profile. In this case, UE power savings can be achieved thanks to the long DRX, and to the facts that a UE in RRC_Inactive can use more relaxed RRM measurements compared to RRC_Connected and performs mobility events (cell reselections) without signalling to the network.

If possible (based on UE support), RRC_Idle mode usage should be limited because of its large CP latency and should be used only for high mobility UEs to avoid mobility related signalling (to RAN/CN).

2.3 Optimized DRX handling for bandwidth adaptation in 5G NR

This section summarizes the ongoing studies oriented to optimize DRX operations for different criteria. In the following we will discuss how to optimize the DRX handling for bandwidth adaptation and low power consumption.

A bandwidth part (BWP) is introduced in NR to support simultaneous operation of UEs with small and large bandwidths in the same carrier. Moreover, different BWPs in a carrier can employ different numerologies. BWP adaptation enables the UE to adjust transceiver bandwidth according to the instantaneous traffic demand so that power consumption can be reduced from the frequency domain. As a result, it serves as an important complementary power saving mechanism to DRX which can further turn on-and-off the UE receiver based on configured timers in response to temporal variation of the actual traffic.

Based on the NR DRX framework, in this section we present three methods for BWP inactivity timer configuration in the context of BWP switching. Specifically, Method 1 aims at having a large BWP inactivity timer value so as to reduce the number of BWP switching operations and the associated signaling overhead. Method 2 is designed to have a small BWP inactivity timer so as to minimize the transmit/receive time of large bandwidth. Moreover, for Method 2, default and non-default BWP based retransmission scheduling schemes are also detailed. In case of non-default BWP based retransmission, one-step and two-step BWP switching approaches are provided. It is up to the gNB to determine which particular BWP inactivity timer configuration method shall be applied in a per UE specific manner. Method 3 gives an optimal BWP inactivity timer determination method for gNB to minimize the overall UE power consumption in the context of bandwidth adaptation. Note that these methods have been detailed in [ONE18-D31], and we summarize them in this section to achieve complete view of the relevant research for this chapter.

Method 1: Robust BWP inactivity timer configuration

In this method, the *bwp-inactivityTimer* should be set to a value not smaller than the sum duration of DL (UL) *HARQ RTT timer* and *drx-RetransmissionTimerDL* (UL). As described in [Section 2.3, ONE18-D31], the *bwp-inactivityTimer* starts or restarts whenever a PDSCH new transmission or retransmission is scheduled in the respective BWP. By configuring *bwp-inactivityTimer* larger than sum duration of *HARQ RTT* timer and *drx-RetransmissionTimer*, the UE shall keep monitoring the PDCCH in the current active BWP before its retransmission timer expires.

Method 2: Energy efficient fast BWP inactivity timer

In this method, *bwp-inactivityTimer* can be set to a value smaller than the sum duration of *HARQ RTT* timer and retransmission timer. In this case, the UE can switch to the default BWP earlier than the retransmission timer expires. If this happens, two scheduling options can be used for the UE to receive the retransmission packet. In one option, the retransmission happens in the default BWP within the retransmission time window, i.e., before retransmission timer expires. This would achieve good energy efficiency by minimizing the unnecessary usage of large bandwidth. In another option, the retransmission takes place in the non-default BWP where the initial transmission happened. In this case, the gNB shall first request the UE to switch to the previous non-default BWP.

Method 3: Efficient BWP inactivity timer configuration

Different options above may lead to different UE energy consumption. The optimal determination of the inactivity timer setting should be configured in a UE specific manner. For cell-center UEs with good SNR and robust MCS, it is envisioned that there is small possibility for retransmission, it may be good to have a smaller BWP inactivity timer value, i.e., Method 2, so as to minimize the usage time of large BWP. But for the cell-edge UE, if the retransmission possibility is large, it can be good to set BWP inactivity timer larger than the HARQ retransmission time window, i.e., Method 1, so as to minimize the signaling for BWP switching. This also depends on the energy consumption comparisons between BWP switching signaling and wideband BWP operation. If the former is smaller, smaller BWP inactivity timer should be configured; otherwise, large BWP inactivity timer can be used.

As a summary of the recommendations from the above presented methods, in fact Method-3 serves as a recommended method, thereby BWP inactivity timer shall be configured according to the UE link quality, e.g., in terms of the retransmission probability so that the overall UE power consumption can be optimized. When UE’s operation bandwidth can timely follow the actual traffic needs so as to minimize the unnecessary large bandwidth usage, the power consumption can be significantly reduced, (e.g., BW reduction from 60 MHz to 20 MHz can decrease ~ 40% power consumption) according to the BW scaling rule of power consumption in Section 2.1.1.

2.4 Signaling and control plane optimizations

This section first presents an innovative network-based device virtualization technique intended for dense environments. The description includes an analysis of the pros and cons of the technique, and proposals for optimizations. Secondly, a detailed analysis regarding optimal RAN split option for the Megacities and Underserved Areas scenarios is performed. This analysis aims to assess the best split option among the list of proposed options from 3GPP that best fits the challenges and needs posed by these particular scenarios, in terms of, e.g., coordination, robustness, infrastructure availability and derived costs.

2.4.1 Network-based device virtualization

A network-based device virtualization technique is proposed throughout this section. By means of this technique, it is foreseen that users can get rid of any actual constraints imposed by the physical device (memory, storage, etc), and feel the impression of a much more capable device since most functionalities are performed by the network.

In the previous deliverable [ONE5G-D3.1], the concept was introduced, exposing the proposed architecture and drafting some immediate benefits of this invention. In order to put the reader in context, Figure 2-6 illustrates the proposed architecture to this end.

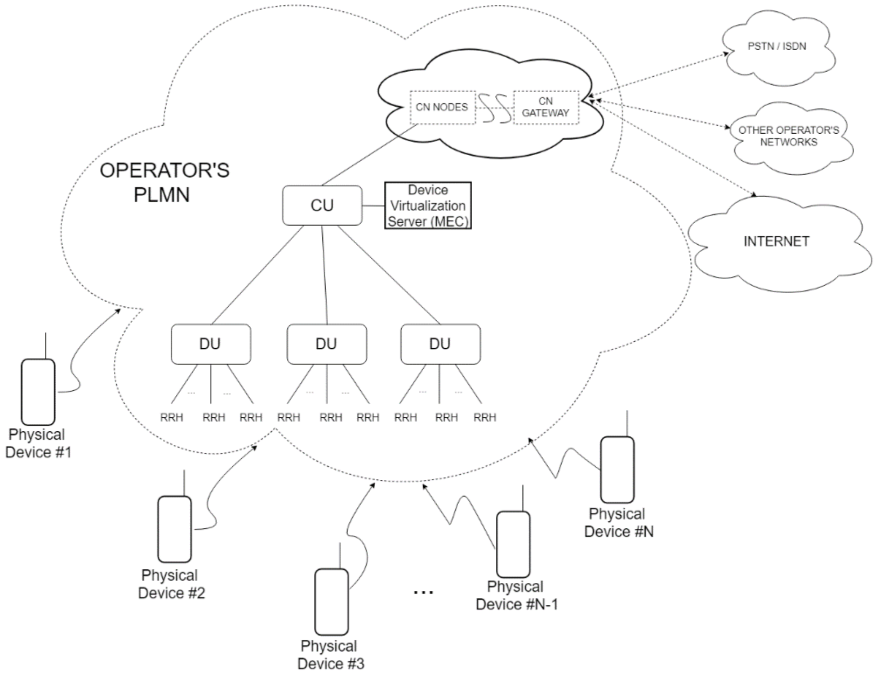


Figure 2-6 Network-based device virtualization architecture

In Figure 2-6, it can be seen that the Device Virtualization Server (DVS) – responsible for delivering virtualized device sessions to the users in the mobile network, and centralizing all connectivity tasks between the users and any other entities in the network – is located at the edge, between the gNB and core network, aggregating a number of RRHs, leveraging MEC functionalities to meet latency requirements, and not affecting QoE.

The main functionalities of the DVS will be:

- To host several applications intended for the users by means of being connected to the operator’s telco cloud.
- To establish a communication path with any peer entity (a remote computer, a terminal device, an Internet-based server, or any other computing device remotely accessible) without any intervention of the physical device.
- To maintain the virtualized device sessions for all the active users and transmit/receive the necessary information to/from the physical devices.

The installation of some demanded applications at the DVS (i.e. at the network edge) and its connection to the operator’s telco cloud allows to guarantee an immediate access to a pre-stablished set of resources, avoiding the users to experience long delays in the communication or feel that content is not being executed locally whenever they make a request. For instance, Virtual Reality/Augmented Reality or gaming applications can benefit from this technique as they can be installed at the DVS in order to reduce E2E latency. Thus, the users would not need to install, upgrade or download any user data and/or application as everything would reside in the cloud and be accessed by the user through the DVS. Devices would only be required to maintain, most of the time, a single low-connection to the DVS that serves the UE’s GUI (Graphical User Interface) to be aware of the actual platforms and services that are being used, and only appropriate delivery of the multimedia content and user interactions are required to/from the physical device. Figure 2-7 illustrates the communication flow between the physical device and the DVS whenever a user makes a request and the connectivity is established. It is to note here that two different concepts are referred to in Figure 2-7 and will appear hereinafter: the “physical device”, which represents the actual device that the user owns and makes use of; and the “virtualized device”, which represents the logical entity created by the DVS node.

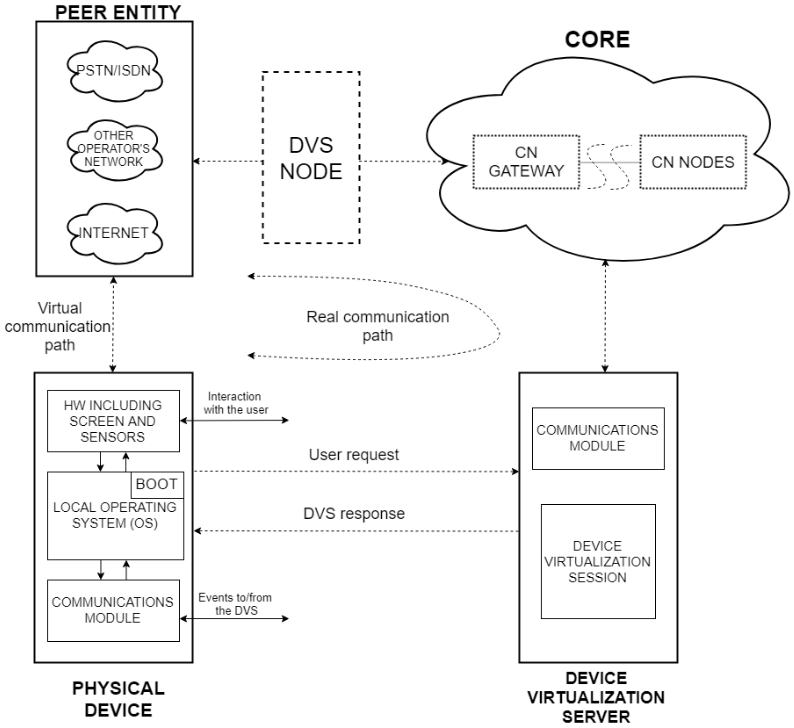


Figure 2-7 Communication between physical UE-DVS

The methodology shown in Figure 2-7 consists of the following steps:

- The user initiates the communication with the DVS by launching a virtualized application in the virtualized device session that is permanently delivered by the DVS based on the list of applications it hosts – User Request.
- The DVS node receives the request from the user and performs all the connectivity actions to establish a connection with the associated peer entity (Internet, PSTN/ISDN, or other operator’s network), which can be located inside or outside of the operator’s network. The communication path with the peer entity may include DVS from another operator, as it is illustrated with a dashed box in the upper part of the Figure 2-7.
- After establishing the connection with the associated peer entity, the DVS will deliver the necessary feedback to the user according to the virtualized device session, whereby the subscriber gets the impression of having a direct “virtual communication path” with the peer side, instead of the real path involving multiple network nodes.

In this context, it can be noticed that the user will benefit from cloud processing as their physical devices would not have to perform any action as all the action will be executed in the Cloud. The DVS node can offer enhanced integrated multimedia services by hiding the complexity of audio, video and data connections to any kind of device regardless of its capabilities and serve the content in a unified way irrespective of the technology involved (circuit-switched, packet-switched, media streaming, and so on).

Therefore, users will only need to be equipped with basic physical devices formed by few elements compared to the traditional devices in wireless mobile networks: an **Operating System (OS)** to locally boot the device and with some basic functionalities to run some applications locally, for those cases where e.g. a fast interaction is needed or the device is out of coverage; **a set of hardware elements**, such as a microphone, buttons, etc., to interact with the network; and **a communication module** to trigger/receive events to/from the DVS. The DVS node will then perform all the connectivity and software-related tasks described above, and the local OS of the physical device will just need to interpret the information received from the DVS and re-create a suitable device session, thus giving the impression to the users that applications are running locally in the device.

The Figure 2-8 illustrates some of the network functionalities that are moved to DVS so as to allow the users to have simpler devices with the same or even higher functionalities.

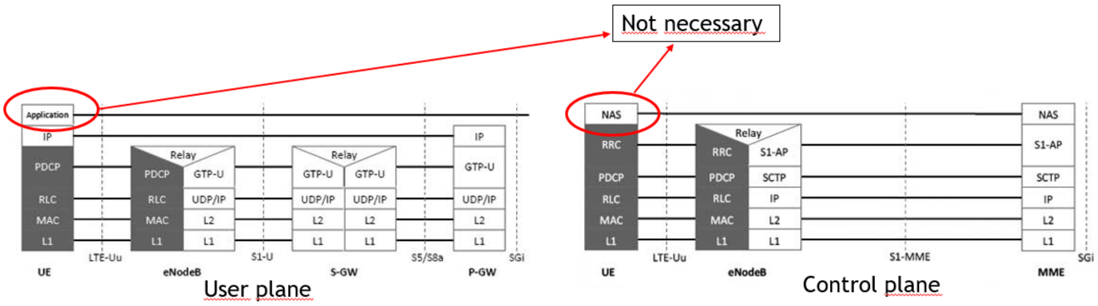


Figure 2-8 DVS impact on user and control plane protocol stack

Network-based device virtualization brings great benefits for users as it allows them to have unlimited access to different services without the need to change their physical devices due to OS (Operating System) incompatibilities, lack of capabilities, etc. Additionally, it gives the possibility of managing all the connections within an area in a centralized way, which are of great interest in closed-spaces such as mega factories.

Nevertheless, UE virtualization imposes challenging requirements to be fulfilled by networks that rely on centralization. For the purpose of this UE device method, it is assumed that CU will have full knowledge about several parameters (e.g. CSI and Resource allocation) strictly required to perform user-cell association and to carry out the process of UE virtualization. However, it should be studied whether those networks can provide such information before going through the centralization.

Among the challenges imposed by this technique, and low-layer centralization itself, the following are highlighted :

- Each UE shall be capable of properly demodulating DL Reference Signals from all transmitting cells comprised in the scenario. Non-orthogonal methods or other feasible techniques should be exploited to this end.
- The length of training signalling under large pathloss considerations from all the RRHs controlled by the DVS should be carefully considered.

In this context, we analyse a multimedia service taking advantage of an upper layer virtualized server (DVS) to optimize the length of training signals under pathloss regime based on compressed sensing (CS). Assuming the existence of an IP multimedia server (IMS) on top of an user IP-CAN session (Application Layer), as depicted in Figure 2-9. In an usual case, an IP-CAN session maintenance procedure will have to reach functions like SMF, AF and PCF regardless the mobility of the UE, the proposal here is to keep alive sessions under the control of the DVS. The CU, where is also placed the DVS, is as well controlling a finite set of RRHs, N_{RRH}, single antenna RRHs whose positions are independently and uniformly distributed over a large deployment area, $\mathcal{A} \subset \mathbb{R}^2$ resulting in a RRH density $\lambda \triangleq N_{RRH}/|\mathcal{A}|$ as in [SW+18c].

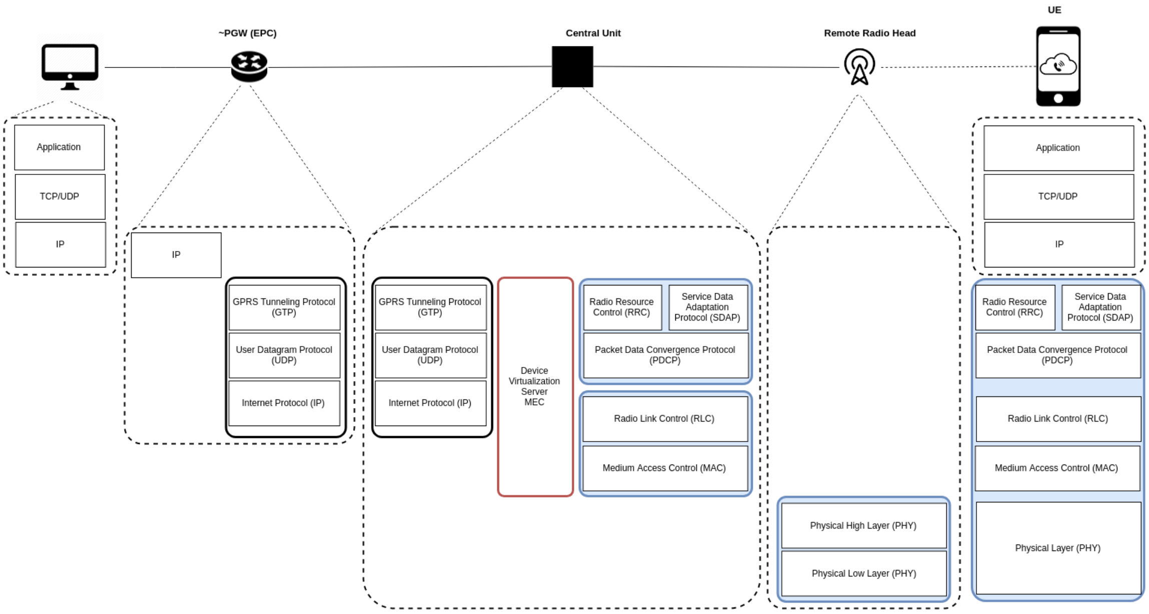

~PGW (EPC)
Central Unit
Remote Radio Head
UE
Application
TCP/UDP
IP
IP
GPRS Tunneling Protocol (GTP)
User Datagram Protocol (UDP)
Internet Protocol (IP)
GPRS Tunneling Protocol (GTP)
User Datagram Protocol (UDP)
Internet Protocol (IP)
Device Virtualization Server MEC
Radio Resource Control (RRC)
Service Data Adaptation Protocol (SDAP)
Packet Data Convergence Protocol (PDCP)
Radio Link Control (RLC)
Medium Access Control (MAC)
Physical High Layer (PHY)
Physical Low Layer (PHY)
Application
TCP/UDP
IP
Radio Resource Control (RRC)
Service Data Adaptation Protocol (SDAP)
Packet Data Convergence Protocol (PDCP)
Radio Link Control (RLC)
Medium Access Control (MAC)
Physical Layer (PHY)

Figure 2-9 DVS Proposed Architecture for Managing RRHs from the CU

Determined by the amount and length of exchanged downlink control signalling, and with the aim of optimize required training sequences, N_p, for computing CSI reports (e.g. CQI) in a dense C-RAN architecture we study the channel state information reference signals (CSI-RS) to adjust a compressive sensing (CS) technique as in [SW+18c]. The CS approach is considered to asses on the analytical limits to reduce training signalling overhead under flat-fading conditions. The design of a CS-motivated training sequence is independent of the number and positions of RRH and UEs. The CSI is considered in the UE side based on CSI-RS, according to a basic numerology frame configuration (e.g. numerology=0). The determined CS bound expression provides insights on how estimation performance is affected by control training sequences length, number of channels to be estimated, RRH density and channel statistics [SW+18c], as in Figure 2-10.

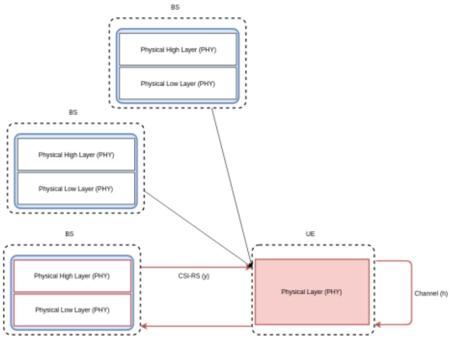
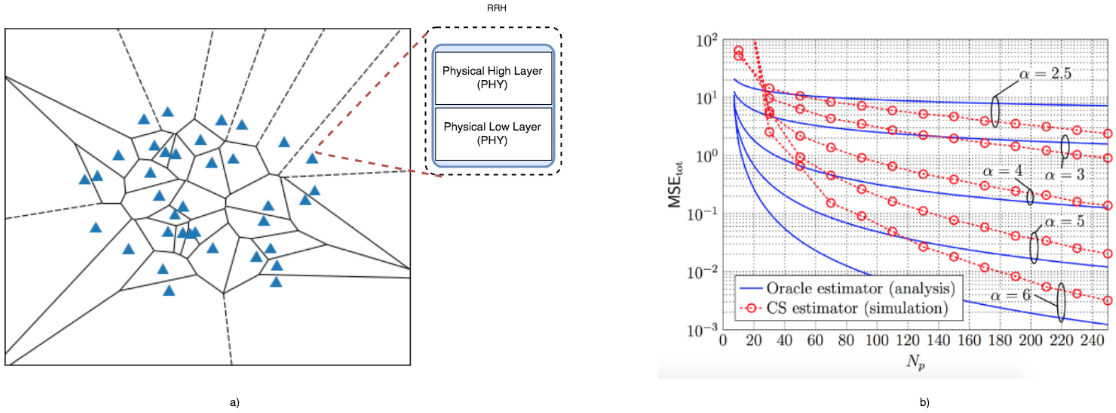


Figure 2-10 RRHs considered on CS

The specific scenario was studied due to the high RRH density under CU control. When a UE has no a priori knowledge of the set of largest-modulus elements of an array of channel values, a CS estimation approach can be employed for estimating the complete channel vector of CSI-RS in a collection of RRH as long as the channel sparsification assumption is valid.

In a DVS/CRAN deployment with $N_{RRH} = 500$ with density $\lambda \triangleq 1$, Figure 2-11 (b) shows the simulated MSE_{tot} performance of the estimator applying the standard Basis Pursuit algorithm on the received signal at the UE end, as a function of the training symbols sequence and for various values of path loss exponents (α), for the system setup considered in Figure 2-11 (a), ($N_{RRH} = 500$). It is seen that the CS estimator performance improves with increasing sequence length N_p, as expected from standard CS theory, and has similar dependence on the pathloss exponent as the oracle estimator, i.e., very good performance is achieved in the large pathloss (α) regime even with small sequence length (N_p) [SW+18c].



a)

b)

Figure 2-11 a) RRH System Setup b) $\mathrm{MSE}_{tot}$for the CS-based channel estimator as a function of the training sequence length Np

In conclusion, an optimal number of estimated channels as in Figure 2-12 is recommended under a DVS/CRAN deployment with a maximum number of RRHs equal to 500. The downlink training sequences shown are equal to $N_p = 50, 100, 150$.

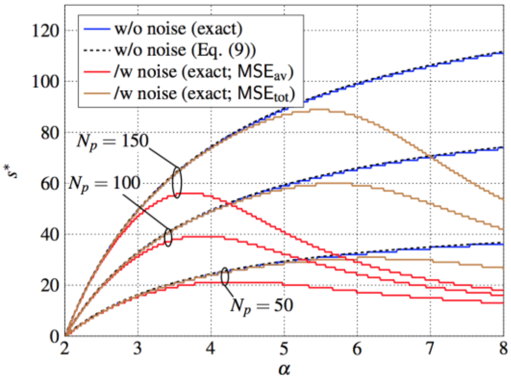


Figure 2-12 Optimal number of channels (s^*) that minimizes MSE as function of pathloss (α)

2.4.2 RAN functional split options analysis and network optimizations for increased robustness

New Radio supports different centralization deployments of the NR radio protocol stack. Thus, different protocol split options between the Central Unit and Distributed Unit (gNB) have been identified to support the different levels of centralization. Figure 2-13 illustrates the different split options properly identified by 3GPP based on E-UTRA protocol stack [3GPP38.801]:

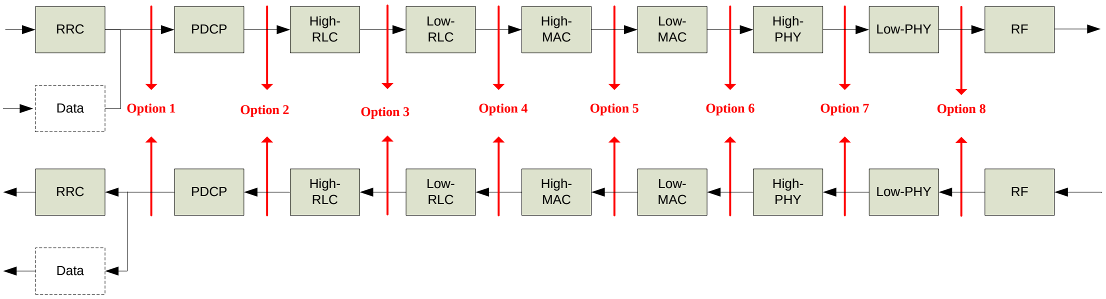


Figure 2-13 3GPP RAN Functional Split [3GPP38.801]

Despite the wide variety of split options showed in Figure 2-13, two high level split options, denoted as ***"High Layer Split option (HLS)"*** and ***"Low Layer Split option (LLS)"***, have been agreed to enable centralization in most of deployments. 3GPP announced the selection of **Option 2 (PDCP)** and **Option 6/7 (MAC/PHY)** as the high layer and low layer split points for split options HLS and LLS, respectively.

New Radio networks shall support the above flexibility on moving some RAN functions to the central unit and/or distributed unit depending on deployment scenarios, the intended services to be supported and their requirements. Nevertheless, the functional split option performed at the gNB would be mostly limited by the capacities of the underlying interface (xHaul).

Following 3GPP terminology, LLS allows the centralization of all high layer processing functions at the CU and enables to implement advanced coordination techniques such as, centralized scheduler or interference mitigation, which results greatly beneficial in high dense scenarios to manage the strong and unmanaged interferences derived from the large number of macro/small cells connected to the single CU.

Nevertheless, LLS requires high performance transport networks, such as dedicated fibre optics or high capacity wireless connections, to support the high capacity data transference that is required in the fronthaul, between the CU and the low-layers of the gNB nodes. In addition, a tight synchronization between the CU and the lower part of the gNB is crucial to perform some of the tasks of the PHY-layer, which results in a very low latency communication between these two entities, e.g. a few hundreds of microseconds.

Moving towards a HLS, the stringent capacity and latency requirements of this split option are relaxed at the expense of reducing the number of processing functions that are centralized at the CU and, in turn, reducing the gain by centralization which can benefit from.

In [3GPP38.801] a study on capacity and latency requirements for different split options is depicted, proposing the values represented in Table 2-6 for HLS and LLS splits:

Table 2-6 System requirements for HLS and LLS options

Split option	**Fronthaul capacity requirement**	**Latency requirement**
High Layer Split (HLS) – Option 2 (PDCP)	DL 4016 Mb/s; UL 3024 Mb/s	1.5 ~10 ms
Low Layer Split (LLS) – Option 6/7 (MAC/PHY)	Opt-6: DL: 10.1~22.2 Gb/s; UL: 16.6~21.6 Gb/s	250 µs

From Table 2-6 it can be seen how the main difference between HLS and LLS relies on the performance of the transport networks for the fronthaul delivery. LLS requires more sophisticated transport networks to deliver the required data rate of the baseband samples (IQ) under stringent latency requirements, as all the computation load is carried out at the CU. In contrast, HLS requires less sophisticated transport networks as most of the processing tasks are performed by the gNB itself. In the latter, the amount of traffic exchanged between the CU and the DU will basically obey to throughput and latency requirements given by the user device, as it occurs in traditionally distributed networks (D-RAN).

Focusing on the scenarios considered in ONE5G project (Megacities and Underserved Areas), it is reasonable that the transport network capacities and operator's infrastructure availability would directly affect the selection of the optimal split option in both scenarios, since a more centralized topology (LLS) is just feasible whether the operator owns dark fiber and can connect to each gNB that the CU is connected to.

Nevertheless, beyond infrastructure availability, there are other factors related to, e.g., system performance or economic impact, that can fairly justify the selection of a certain split option despite the lack of a technological justification. In that case, a trade-off between system performance and technology performance will have to be found to outcome the most appropriate split option that ensures an optimal balance for that situation. In any case, the selection of a certain split option is not exclusive since other split option can also be valid. Likewise, a split option selection may vary among similar areas sharing the same commonalities, with regard to the type of scenario or services supported, due to technical or commercial factors which are market-and operator-specific.

Going through the aforementioned scenarios, the main difference between megacities and underserved areas is network density. **Megacities** are characterized by a large number of cells separated hundreds of meters between them, operating at high and low frequencies and/or implementing advanced features in order to cope with the high traffic density demands in overpopulated areas. In contrast, **Underserved Areas** are characterized by pursuing large coverage areas instead of high data rates by means of deploying few cells located far away from each other with the aim of covering several regions.

Hence, the objective pursued for each scenario varies significantly and, in turn, the requirements set on the system performance. In the case of **Megacities**, where a huge interference is produced due to the large number of cells, a cross coordination of cell transmissions would be highly beneficial by means of a **LLS centralized topology** where the CU can efficiently and dynamically manage harsh interference transmissions, for instance, by implementing Joint-Transmission CoMP.

As an example, in a C-RAN deployment with LLS, the CU would have full knowledge of channel conditions of each pair User-gNB (as everything up to PHY layer is centralized) and it can choose, in every timeframe, the bandwidth portions where no huge interferences of neighbour cells are detected to allocate certain cell transmissions. Further, the CU can implement coordination techniques to create a much more complex cooperative scenario among the cells deployed, i.e. in a MIMO-fashion network.

Besides the benefits derived from a cooperative processing and coordinated scheduler by enabling the CU orchestrating all cell transmissions, Megacities can also benefit from hardware costs savings by performing a low-layer split.

The centralization of almost all RAN functions at the CU allows simplifying gNB architecture by reducing the amount of hardware equipment needed to be installed at the site. Therefore, it is foreseen that the deployment costs derived from an LLS centralized network will decrease as just few hardware components are now required to be installed at the site and everything resides in the CU. For instance, one of the most important hardware reduction savings would come from leased rooftops, as less equipment is needed to be installed at the rooftop, site rental costs would be reduced.

Furthermore, operational and maintenance expenses seem to experience a similar reduction as well. With LLS, a better failure detection can be made through the CU enabling taking quick actions and changes over the network, which would significantly reduce the outage probability. As less outage time is foreseen to occur during operation, less maintenance actions would be required to be depicted on the site, which results in greater savings.

Unlike Megacities, **Underserved Areas** cope with other challenges related to the type of environment, i.e. topography and accessibility. In general, the accessibility in these areas is quite difficult and impractical and it's a real challenge to deploy a telecom infrastructure to provide basic connectivity for the users. Concretely, the main problem derived from these areas is at the backhaul network. In most cases, there is no operator's infrastructure available in these areas and the deployment of a new infrastructure may show to be unreasonable due to the high complexity and expenses.

Renting a satellite segment to handle backhaul traffic is the most appropriate solution to cope with the unfeasibility of deploying a terrestrial network in these remote areas, depending on whether the type of service to be provided does not have any latency constraints. However, this solution incurs more expenses compared to traditional terrestrial networks such as fibre optics or microwaves, which may result not affordable when studying the business case of these areas.

In addition, these areas are certainly vulnerable given the probability of a natural disaster, or a catastrophe, as well as the harsh environment such as high/low temperatures, snow, animals' actions, etc., affecting the communication established with the users and preventing a basic connectivity for the users that are located within these zones.

In such a way, the objective pursued by these areas relies on investigating network optimizations that allow ensuring a basic connectivity for users that are placed in Underserved/Inaccessible Areas and can also serve to reduce the incurred expenses. Thus, a centralized topology deployed through a **High Layer Split (HLS)** can result beneficial for these remote areas in order to minimize the backhaul access and increase robustness against the above sources of disasters.

By means of centralization, the CU will be able to gather the traffic coming from all the cells that are under CU umbrella. Thus, implementing centralization in Undeserved/Inaccessible areas allows aggregating all data traffic coming from a certain area –several remote cells- at the CU and, in turn, avoiding continuous access to the backhaul network, which certainly will be cost-efficient. Nevertheless, in these areas a tight coordination and cooperation among the cells aggregated by the CU is not required, since strong interferences are not foreseen. Therefore, a high layer split option would be the preferable option.

Additionally, through the CU it would be feasible to perform local breakout of data traffic going towards Internet in order to handle it locally and, thus, improve efficiency. The CU would have then twofold functionality in these areas: serve as a traffic aggregator and perform local breakout. The latter can be achieved by implementing MEC technology at the CU by means of a server equipped with MEC technology co-located with the CU.

Coming to an end with the above analysis, a **Low Layer Split (Option 7)** seems to be the most suitable level of centralization for **Megacities scenarios** in order to create a cooperative and collaborative scenario dealing with harsh interferences that impact negatively the user experience. On the contrary, a **High Layer Split (Option 2, or even Option 1)** seems to be more suitable for **Underserved Areas** where a low grade of aggregation allows to reduce total costs derived from expensive backhaul access through, for example, satellite. Nevertheless, a detailed cost analysis should be performed before choosing one of the above configurations, as most likely there should be a trade-off between the benefits and the total cost.

Despite the multiple gains apparently yielded from choosing a LLS centralized architecture for a Megacities scenario or a HLS centralized architecture for Undeserved Areas, it would be crucial to ensure that the transport networks will meet roughly the requirements shown in Table 2-6, as to avoid bottlenecks and guarantee quality of communication.

2.5 Summary

Table 2-7 Summary of key recommendations and benefits in terms of control signaling and UE power consumption

Feature	Recommendation	E2E / KQI benefits
Service dependent RRC state handling	RRC state handling to be optimized for latency, UE power and network signalling based on service/traffic requirements as well as mobility profile of a UE. Key recommendations: RRC_Connected with long DRX to be used at medium/high traffic for semi-stationary UEs. RRC_Inactive with longer DRX to be used for infrequent traffic and low/medium mobility profile. RRC_Idle mode usage to be limited only at high mobility.	It achieves reduced network/service accessibility (i.e. shorter control-plane, CP, latency), service retainability (i.e. longer battery life), and better service integrity (overall lower user-plane latency): Up to 89% shorter CP latency at transition from RRC_Inactive to Connected (compared to Idle). ~70% / ~40% longer battery life of RRC_Inactive compared to Connected in no data scenarios & infrequent traffic, respectively. Only ~10% latency increase of RRC_Inactive compared to Connected for infrequent traffic.
Efficient BWP inactivity timer configuration	BWP inactivity timer shall be configured according to the UE link quality, e.g., in terms of the retransmission probability so that the overall UE power consumption can be optimized.	Improved UE power efficiency. Specifically, when UE's operation bandwidth can timely follow the actual traffic needs so as to minimize the unnecessary large bandwidth usage, the power consumption can be significantly reduced, (e.g., BW reduction from 60 MHz to 20 MHz can decrease ~ 40% power consumption) according to the BW scaling rule of power consumption in Section 2.1.1. .
Device Virtualization	DVS introduction allows to host some of device functionalities (less protocols in SW stack) and abstract users from CN interaction. To support DVS introduction in a CRAN deployment a maximum number of RRHs = 500 is recommended, as well as, a CS approach for CSI estimation under pathloss and noise considerations.	Improved user QoE and simplify user device. Improved flexibility and adaptability to different requirements. ~ 87% overhead length reduction for downlink training sequences under a dense DVS/CRAN deployment is achieved for different pathloss and number of estimated channels.

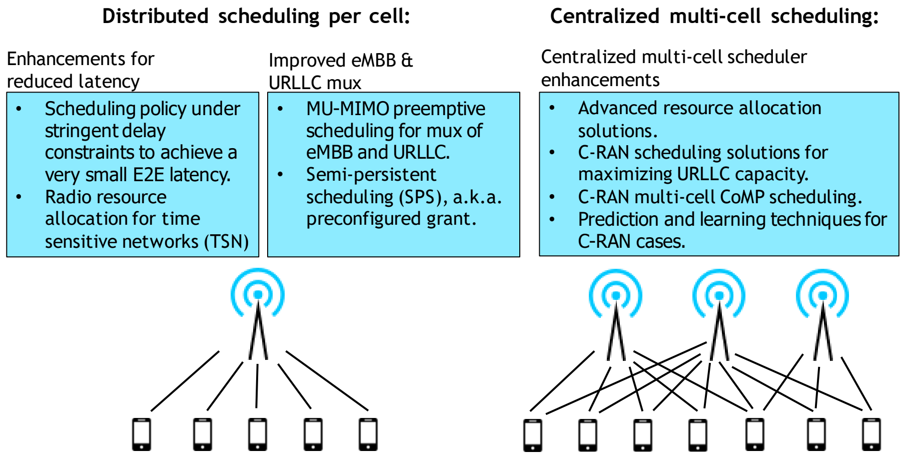
	Optimal number of estimated channels is presented for a dense scenario comprising 500 RRHs orchestrated by a DVS/CU element.	
Cloud RAN Split options	Analyse of relevant split options tailored to the type of scenario (megacities and undeserved areas). The Low Layer Split (Option 7) is most suitable for Megacities scenarios. The High Layer Split (Option 2, or even Option 1) is more suitable for Underserved Areas.	Improved network availability Improved service retainability Improved transport network utilization

3 Multi-service and context aware radio resource management optimization

This chapter addresses the problems of resource allocation to fulfil end-users QoS requirements, including multi-service use cases, and exploiting elements of context awareness. Section 3.1 presents various advancements in MAC-level scheduling, ranging from distributed single-cell scheduling to centralized multi-cell scheduling. Section 3.2 presents derived performance optimization schemes for multi-node and/or carrier aggregation use cases.

3.1 Multi-service scheduling solutions

In this section, we present a set of promising multi-service scheduling solutions that help further enhance the E2E performance, mainly by improving on the KQI service integrity. Figure 3-1 shows an overview of the three main categories of scheduler enhancements. As pictured, generic MAC scheduling policies to improve the latency are proposed for distributed architectures, where scheduling is performed independently per cell. This includes also enabling the new 3GPP NR Rel-16 use case that relates to Time Sensitive Networks (TSN). Moreover, enhancements to facilitate more efficient multiplexing of highly diverse services with significant QoS targets – such as eMBB and URLLC service categories – are proposed. Another proposal to multiplex low-latency traffic with broadband traffic in a flexible Time Division Duplex (TDD) system is introduced in Section 7.6 (Appendix F). Third, a set of promising multi-cell scheduler enhancements for advanced C-RAN architectures are developed to fully unleash the potential of these architectures. We focus on a C-RAN architecture with the lower layer split, where the MAC (hosting the scheduling functionality) is located in the centralized unit (CU). Notice that due to the user-centric design of the 5G NR, the developed enhancements are largely complementary to each other, and can to a certain degree be applied in combination. Although the enhancements developed specifically for C-RAN architectures of course are not directly applicable for D-RAN architectures.



Distributed scheduling per cell:

Enhancements for reduced latency

- Scheduling policy under stringent delay constraints to achieve a very small E2E latency.
- Radio resource allocation for time sensitive networks (TSN)

Improved eMBB & URLLC mux

- MU-MIMO preemptive scheduling for mux of eMBB and URLLC.
- Semi-persistent scheduling (SPS), a.k.a. preconfigured grant.

Centralized multi-cell scheduling:

Centralized multi-cell scheduler enhancements

- Advanced resource allocation solutions.
- C-RAN scheduling solutions for maximizing URLLC capacity.
- C-RAN multi-cell CoMP scheduling.
- Prediction and learning techniques for C-RAN cases.

Figure 3-1 Overview of proposed scheduler enhancements

3.1.1 Spatial and temporal availability of URLLC services

Among the established 5G NR service categories, URLLC will pave the way for numerous verticals, such as Internet of Things (IoT), mission critical services, real-time control and automation for multiple market segments. However, URLLC is accompanied by a plethora of challenges and stringent requirements to ensure service operation with virtually no failures during the operation time. These challenges are currently being addressed by 3GPP and different industry associations/ standardization bodies [3GPP22.261]. URLLC service performance has been mainly evaluated by means of metrics such as packet error ratio, latency and jitter. These metrics, though fundamentally meaningful from the radio communication perspective, need to be looked collectively with the service demands from a vertical-specific point of view (e.g., availability of a service and reliability of its operation).

Concentrating on both space and time domains, in this contribution, we make a first attempt to *bridge the gap between traditional radio-link KPIs and service-level KPIs* by providing an insight on the availability and reliability of wireless links in 5G systems. We propose the definition of a new, stochastic quantity to measure the spatial availability of a given service characterized by a QoS requirement subject to a service-specific confidence level, when communication takes place via a wireless link between a given End Point (EP) and its serving Access Point (AP). Moreover, capitalizing on the proposed definition of spatial availability, we propose a novel resource allocation scheme dependent on the spatially available area of a given AP and based on the concept of resource provisioning.

3.1.1.1 Spatial availability of a service

Spatially, a novel, service-related definition of spatial availability, based on the signal quality (i.e., Signal-to-Interference Ratio - SIR) is introduced [EFS18], taking into account the confidence level of overcoming a predefined SIR threshold. We present the following definition of *(θ, α)-availability*:

Definition: Any EP located at point $\boldsymbol{i}$ of a 2D plane and served by an AP located at point $\boldsymbol{j}$ is labeled as (θ, α)-available, if $\Omega_{\boldsymbol{i},\boldsymbol{j}} = 1\left(\Pr\left[\mathrm{SIR}_{\boldsymbol{i},\boldsymbol{j}} \geq \theta\right] \geq \alpha\right) = 1$; $\boldsymbol{i},\boldsymbol{j} \in \mathbb{R}^2$; and non-available otherwise.

Given this definition, for a generic AP, the ratio of its (θ, α)-available area (i.e., the collection of all (θ, α)-available EPs) over its overall geographic area (i.e., Voronoi region) is denoted as the spatial availability (ratio) of the AP, termed hereafter as A_s. In Figure 3-2, a deployment visualization of three different (θ, α)-available areas is shown, where the different colors represent different QoS targets, i.e., different values of targeted SIR, θ, and/or the confidence level, α.

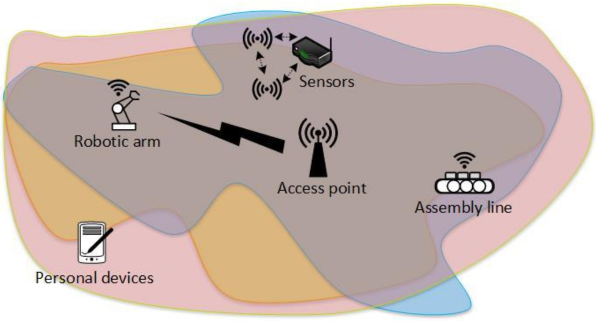


Figure 3-2 Qos-available areas for a generic AP with different EPs deployed. Different coloured regions refer to different QoS requirements (e.g. signal quality measured by means of SIR and/or probability to achieve a targeted QoS level)

To examine the effect of having different SIR thresholds and confidence levels to meet such a threshold, numerical evaluations based on spatial averaging were conducted for a deployment area of $100\text{x}100\ m^2$ for a total of 10 APs. For more details regarding the simulation assumptions the reader is referred to [EFK18]. The spatial averaging is adopted to ensure deployment-agnostic results, which help in reaching a system-wide insight about the behaviour of the spatial availability for different system parameter values. Figure 3-3 shows the effect of varying requirements (i.e., SIR threshold, confidence level and APs' spatial intensity) on the spatial availability. On the left figure one can observe an expected decrease of the spatial availability as the requirements become more demanding (that is, as the values of θ and/or α increase). On the right figure, we observe only a slight gain in the spatial availability as the number of deployed APs over a fixed area increases. Such a behaviour is justified as follows: on one hand, for a given EP, as the number of deployed APs increases, the received useful signal power also increases due to the increased proximity between the EP and its serving AP, nevertheless, the interference from neighbouring APs increases as well, hence, resulting to a slow degradation of the received SIR, which, in its turn results to a slow decrease of the (θ, α)-available area. On the other hand, the Voronoi region of the focused AP decreases as well, however with a slightly higher rate, thus, leading to a slight increase of the spatial availability. These observations, thus, justify the call for efficient interference management techniques among the APs to drastically enhance the spatial availability.

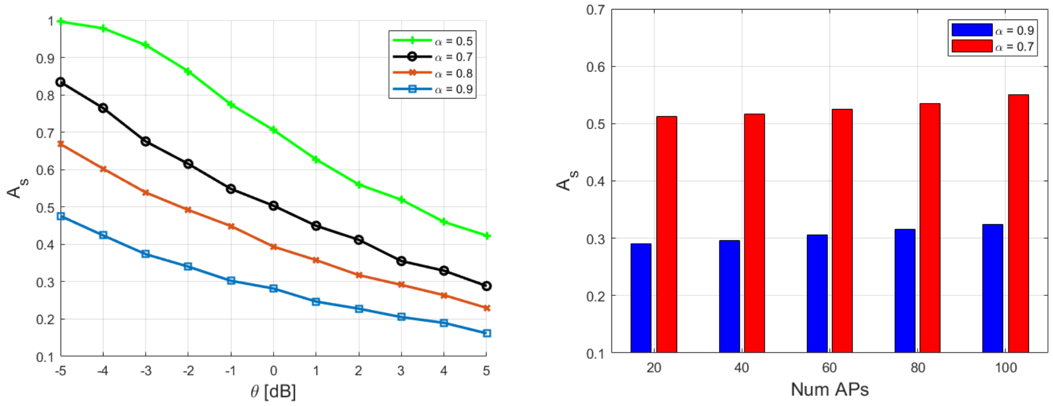


Figure 3-3 (left) Effect of different requirements on the achieved spatial availability (right) Effect of network densification

3.1.1.2 Temporal availability of a service

Based on the proposed spatial availability metric, and since our objective is to propose a unified, space-time availability framework useful to evaluate the impact of resource provisioning for URLLC systems, a *spatial availability-proportional resource (i.e., channel) reservation scheme* is also proposed. Each EP is considered to have service requests arriving temporally based on an exponential distribution, whereas all service requests are assumed uniformly distributed over space. Such requests need to be addressed by the serving AP which can exploit a specific number of available resources.

In further detail, each AP is entitled to use M available channels that can be accessible by the EPs within its geographical coverage (i.e., Voronoi region) within a frequency hopping framework. Since the earlier defined spatial availability A_s decomposes the Voronoi region into two disjoint (i.e., spatially available/ unavailable) regions, the number of resources to be utilized by EPs located in the (θ, α) - available and non-available regions of this AP can also be decomposed proportionally to the spatial availability A_s.

The goal of this resource allocation scheme is also to investigate the temporal "analogous" of (θ, α)-spatial availability, which is called the system's *temporal availability* and is defined as the probability of having at *least one free resource to be utilized by an incoming EP request at a specific time instant*. More information regarding the analytical tools on how to derive the temporal availability and its relation to the temporal reliability of a wireless service can be found in [EFS18]. In Figure 3-4, the temporal system performance is shown, focusing on EPs located within the (θ, α)-available area of an AP. The results clearly show that (i) time reliability is upper bounded by time availability (as expected – left hand side figure), as well as (ii) the effect of the arrival-to-service rate ratio for EPs under the coverage of the given AP (right hand side figure). With regards to the right hand side of the figure, it should be noted that the steady state time availability is a time-independent metric and can be interpreted as the average operating (or, "up") time of the system. Such performance results can, therefore, provide deeper insights on designing almost deterministically performing wireless systems.

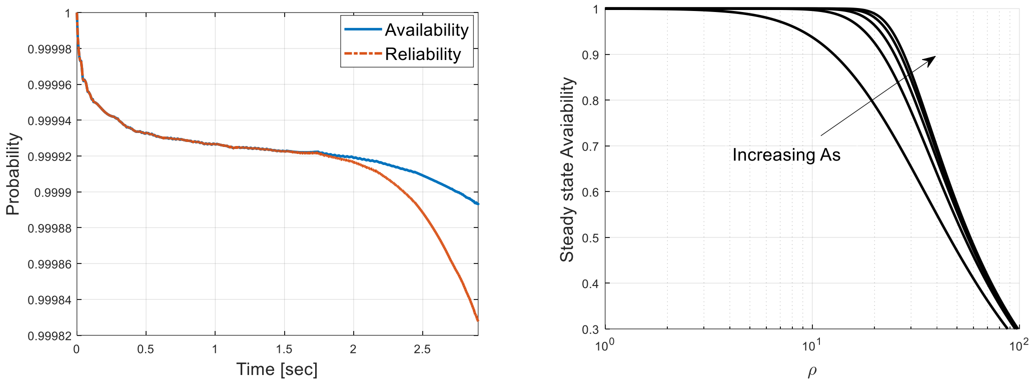


Figure 3-4 (Left) Temporal availability and reliability for A_s=0.8 and M=30 for the (θ,α)-available area (Right) Steady state time availability as a function of ρ (arrival/ service rate ratio of EPs) for increasing values of A_s

3.1.1.3 Joint space-time analysis – the vitality of resource provisioning

Finally, a joint analysis of spatial and temporal availability can give key insights on guaranteed performance for a given wireless link, which is essential for URLLC services. In that sense, the relationship between the steady state time availability and the spatial availability is illustrated in Figure 3-5 for different values of M. First, we observe a symmetric time availability performance for a fixed M between the (θ,α)-available and non-available areas, due to the proposed channel allocation scheme. For As = 0.5, the number of channels allocated to each area will be identical, hence, leading to an identical time availability performance, as requests arrive uniformly in space. Additionally, fixing the value of A_s, time availability increases together with M. This result intuitively emphasizes the role of *redundancy* and *provisioning* in wireless systems.

Equivalently, through our proposed space-time analysis, the minimum total number of channels needed to achieve a targeted temporal availability level can be identified. To further highlight this, a steady state time availability requirement of 0.8 is marked for $M = 20$ and $M = 30$ curves. As expected, the range of A_s meeting the imposed requirement is larger in the latter case. This means that, resources provisioning provides additional redundancy to the time availability performance of multiple service classes.

In conclusion, a unified framework of computing temporal and spatial availability for a wireless-based system was presented. A novel, service-driven definition of spatial availability was introduced, taking into account the probability to achieve a targeted SIR threshold with a given confidence level. Temporal availability was investigated considering a novel, space availability-proportional channel access scheme based on the concept of channel provisioning, bringing up the coupled relation between spatial and temporal availability and reliability. The study is supported by numerical evaluation results which underline the impact of different system parameter values on space/time availability and time reliability, as well as the coupled nature of these performance metrics.

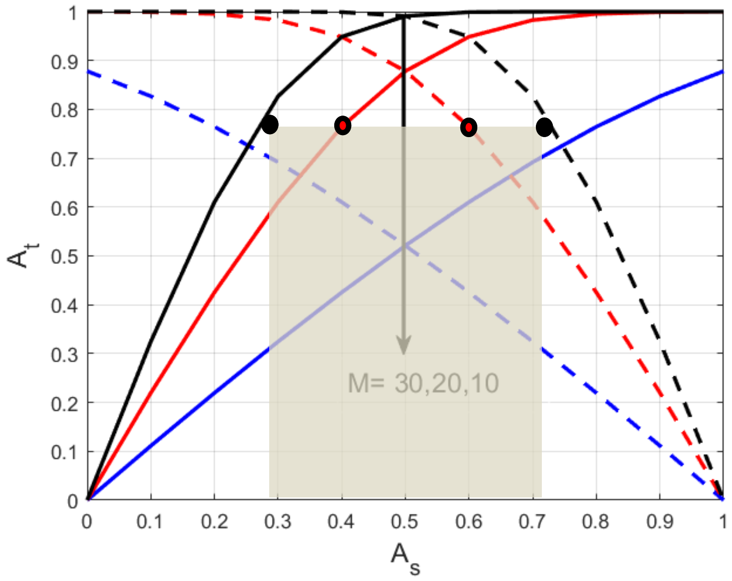


Figure 3-5 Steady state time availability as a function of spatial availability for varying numbers of channels (M = (10, 20, 30)). Dashed and solid lines represent (θ,α) - available and non-available areas respectively

3.1.2 Scheduling of latency critical traffic

This subsection provides a description of two scheduling policies that aim at allocating the RBs among the users in such a way to improve the service integrity for eMBB and URLLC traffic (KQI-I and KQI-sI). In the first subsection, we focus on eMBB services with the objective to minimize the average packet delay in the network, while in the second subsection the focus is on URLLC traffic with the goal to improve the E2E reliability, latency and throughput.

3.1.2.1 Delay optimal multi-user scheduling

We consider in this section the general problem of multi-user scheduling in 5G where the gNB has to allocate the Resource Blocks (RB) among the users at each timeslot or TTI (Transmit time Interval). The system model considered in this section is compatible with the 5G NR Release 15 and can be used for any frame size. The main focus of the work is to schedule the users in such a way to minimize the total average delay under the assumption of bursty traffic arrival (assumed in this section to be generated according to a uniform distribution). This class of scheduling policies is called delay optimal scheduling. It is worth mentioning that the work described in this section is more suitable for eMBB services, as it focuses on average delay minimization. Furthermore, minimizing the average delay will improve the service integrity (KQI-I) for various eMBB service categories (e.g. file transfer delay, page download time, etc.).

In more details, we consider a wireless network composed of one base station serving N users where N could be large. Furthermore, we assume that there are M resource blocks to share among the users at each timeslot. At the BS, there are N queues (i.e. a queue for each user) in which packets are buffered before being transmitted to the users. In addition, we consider that the transmission rate to each user is at most R_i packets per timeslot, where i is the user index. In order to improve the QoS of the users, we are interested in this section in finding the scheduling policy that minimizes the long-run average delay of users' packets. Even though at a first look this scheduling problem seems to be a simple one, we show that it can be cast as a Restless Bandit Problem (RBP), which can be seen as a particular case of Markov Decision Processes (MDP)However, RBPs are PSPACE-Hard, which means that their optimal solution is out of reach. One should therefore propose sub-optimal policies when dealing with such problems. The main interest of the work in this section is that we propose a scheduling policy that provides performance (i.e. total average delay) very close to that obtained by the optimal one in some cases. The approach is based on using the Lagrangian relaxation technique. By observing that the Lagrangian expression has a separable structure, we decompose it into much simpler one-dimensional relaxed problems that can be analysed more easily than the original one. The solution of the individual relaxed problems allows us to develop a simple scheduling policy. This policy consists in computing, at each timeslot, a given index for each user and to schedule simply the users with the highest indices. Interestingly, we proved mathematically that our scheduling policy is asymptotically optimal in some cases. For instance, if a user can be allocated at most one RB at each timeslot, the proposed policy is optimal when the number of users is large. Although this assumption may appear restrictive, it may arise in practice when the number of users is very high as in this case it is more likely that a given user will not be allocated more than one RB at a time. For more details on the development of the policy, the computation of the indices of the users and the proof of asymptotic optimality, one can refer to [KLA18].

To corroborate the aforementioned theoretical analysis, we provide in Figure 3-6 a comparison in terms of average queue length (which, by Little’s law can be translated into average packet delay) between our policy (denoted by WI), the optimal solution and the max weight policy. The max weight policy consists of allocating the resource to the user having the maximum weighted rate where the weight is the queue length. We consider a simple setting of two classes of users with maximum rates given respectively by $R_1 = 5$ Mbps and $R_2 = 10$ Mbps and a total number of users equal to $N = 2M$. The results in Figure 3-6 show that our policy gives better performance than the max weight policy (MW) and it is near optimal when the number of users is very large. The main conclusion is that some myopic scheduling policies (e.g. max weight) that ignore the dynamic traffic arrival are suboptimal from an average delay point of view. In addition, we compare between our policy and Proportional Fairness (PF) scheduling for 100 RBs and various numbers of users. The results, depicted in Figure 3-7, show that the gain (in terms of average queue length) obtained by our policy is high especially when the number of users is higher than the number of RBs. The main conclusion is that some myopic scheduling policies (e.g. max weight or PF) that ignore the dynamic traffic arrival are suboptimal from an average delay point of view. Our policy is asymptotically optimal in some scenarios as explained above.

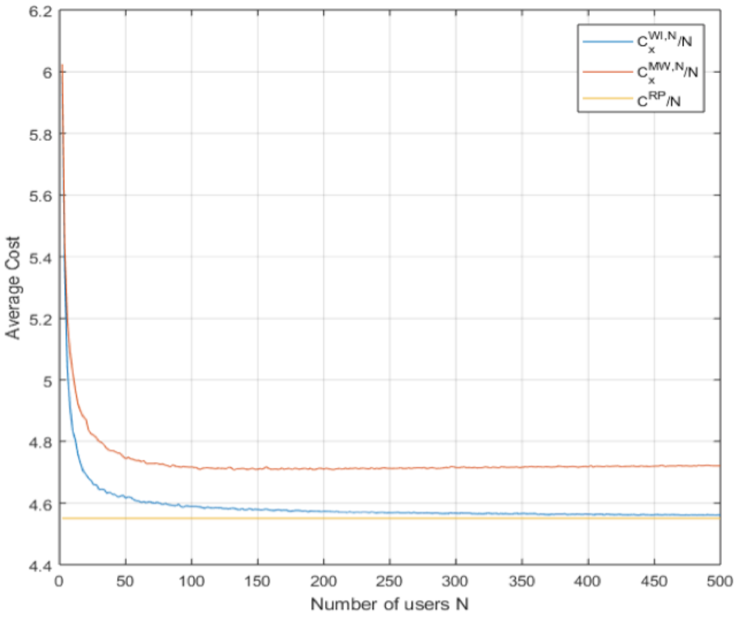


Figure 3-6 Average queue length (cost) vs number of users

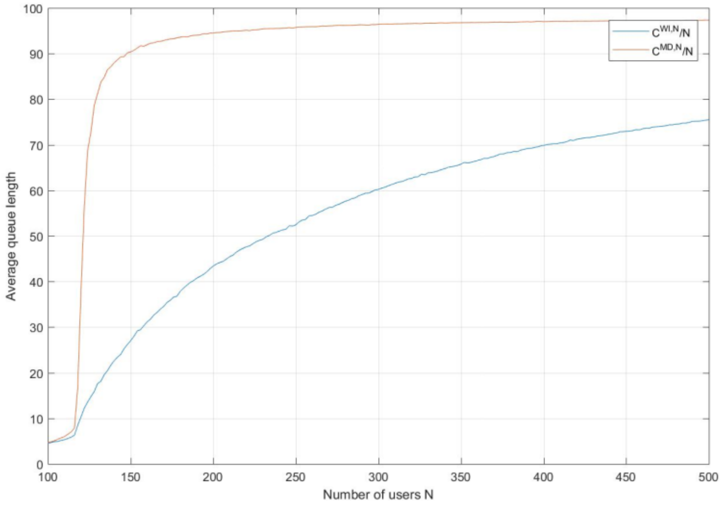


Figure 3-7 Average queue length (packets) vs number of users

3.1.2.2 Dynamic resource allocation for URLLC services

We study the problem of RB allocation for URLLC traffic in a single cell multi-user multi-channel wireless network where URLLC packets have to be transmitted over time-varying and fading channels. Due to the strict latency requirement of URLLC packets, there is not enough time for the BS to estimate the Channel State Information (CSI), and the packets are transmitted in the absence of CSI at the transmitter side. When a packet is successfully decoded, the receiver sends an acknowledgment feedback, which is assumed to be instantaneous and error-free. We consider a discrete-time system where a centralized controller (which is the BS) decides at each time slot on the number of parallel transmissions (or equivalently, the number of resource blocks) to allocate to each user based on his perceived QoS in the previous time slots.

More precisely, for each user k, we define the following parameter ρ_k:

$$\rho_k(t) = \frac{1}{t}\sum_{i=0}^{t-1}\frac{n_k(i)}{a_k(i)},$$

where $n_k(i)$ denotes the number of unsuccessfully decoded (dropped) packets for user k at time step i, and $a_k(i)$ denotes the number of arrived packets at the gNB (for user k) at time step i. The parameter ρ_k reflects the perceived QoS of each user, that is the packet loss rate, and it evolves in a stochastic manner depending on the number of parallel transmissions allocated to that user, the wireless channel, and the arrival traffic. We aim to find an efficient dynamic scheme to allocate the RBs to the users such that the parameter ρ_k for all the users remains within an acceptable range. Therefore, we consider the framework of Markov Decision Processes (MDP) to solve the decision problem.

We define a finite-horizon MDP over the interval [0,..,T], wherein the state space is composed of the Cartesian product of ρ_k over all the users $1,..,K$; and the reward is the expected total number of successfully decoded packets. We search for a policy, π, that solves the following optimization problem.

$$\begin{cases} \max_{\pi} \sum_{t=0}^{T} \mathbb{E}^{\pi}\left[r_i(s_i, \ell_i)\right] \\ s.t.\ \ \mathbb{P}(\rho_k(t) \geq \rho_{max}) < w, \\ \forall\, t\ \in [0,..,T]\ \text{and}\ k\ \in [1,..K], \end{cases} \qquad \text{(P1)}$$

where $r_i, s_i,$ and ℓ_i denote respectively the reward (throughput), the system state, and the action at time slot (or decision epoch) i (recall that a policy is a sequence of actions at different time slots). The symbols $\mathbb{E}$ and $\mathbb{P}$ stand for the expectation operation and the event probability, respectively.

In other words, we look for a policy that maximizes the total expected reward over the planning horizon, while keeping the probability that ρ_k exceeds a threshold value ρ_{max} for any user smaller than a fixed value w.

In MDP theory, classical methods such as value iteration, policy iteration, or linear programming are usually used to find the optimal policy with respect to a certain criterion [Put05]. However, these exact methods require a perfect knowledge of all the model parameters, and hence cannot be applied in our context since we make the assumption that the BS transmits packets in the absence of channel statistics. In fact, we consider a scenario where the BS has neither the instantaneous CSI nor the channel statistics.

To find the optimal policy for problem (P1), we resort to a reinforcement learning algorithm (Q-learning) [SB98]. This method enables the controller (which is the BS in our context) to learn the optimal policy from the experience and in the absence of model parameters. During the learning phase, the controller gets estimates of the value of each state-action pair. It updates its estimates through the interaction with the environment where, at each iteration, it performs an action and then observes both the reward and the next state.

Due to the problem of a very large state space, we conducted numerical studies in the scenario of two users. In the following, we denote by Π^* the maximum-value policy, that is the policy which maximizes the total expected reward, and by $\overline{\Pi}^*$ the minimum-risk policy, that is the policy which minimizes the event probability given in (P1). Besides, we define Π_f as the static or fixed policy which consists in allocating the same number of RBs to the users at each time slot, and finally Π_ξ^* denotes the weighted policy that maximizes a weighted sum of both criteria in (P1).

We depict in Figure 3-8 (left), the total expected reward over the time interval $[t, T]$ denoted by $u_t(s)$ as a function of time when the initial system state s =(0.3, 0) and different policies are followed (recall that the system state is defined above as the Cartesian product of ρ_k). We observe that the maximum-value policy Π^* clearly outperforms the other policies. In Figure 3-8 (right), we display the probability of violating the QoS requirement (event probability in (P1)) over the interval $[t, T]$, denoted by $\bar{u}_t(s)$, when different policies are followed. We observe that the performance of the minimum-risk policy $\overline{\Pi}^*$ is much higher than that of the other policies, (namely the fixed policy Π_f) as it ensures the lowest QoS violation probability ($\bar{u}_t(s)$). For example, at time step $t = 5$, $\bar{u}_t(s)$ is equal to 0.02 when the policy $\overline{\Pi}^*$ is performed whereas this value amounts to 0.42 when the policy Π_f is followed.

Note that $\bar{u}_t(s)$ decreases over time, because as time goes on, the remaining time interval decreases (recall that we consider a finite-horizon MDP), and hence the probability of QoS violation over the remaining time slots decreases. However, $u_t(s)$ increases until reaching a maximum.

Our main conclusion is that, in the cases where the CSIs are absent at the transmitter side, the BS has to allocate multiple RBs to each user to increase the communication reliability Instead of allocating a fixed number of RBs, we have shown that the system performance can be substantially enhanced, by adapting dynamically the number of RBs to allocate to URLLC users (i.e. the number of parallel transmissions of a packet must change over time). For more details, please refer to [BAD19].

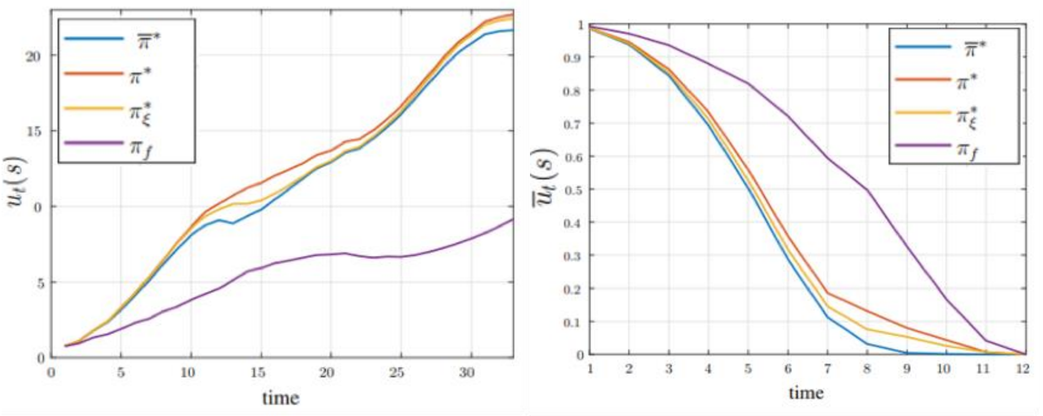


Figure 3-8 (Left) Reward (Expected number of successfully decoded packets) over time. (Right) Probability of violating the QoS requirement over time when for different policies: the minimum-risk policy $\overline{\pi}^*$, the maximum value policy π^*, the weighted policy π^*_ξ, and the static policy π_f

3.1.2.3 Configured grants for periodic non-synchronous uplink URLLC traffic

In 3GPP release 15, semi-persistent scheduling (SPS) can be used to allocate resources to uplink transmissions for URLLC traffic without requiring the UE to request a grant for each transmission. Specifically, in 3GPP TR38.825 ”Study on NR Industrial Internet of Things (IoT)” the SPS or configured grant approach is suggested for TSNs when traffic periodicities are misaligned and cannot be scheduled periodically according to the expected arrival times. Such misalignment typically arises from clock time deviation in cases where UEs are not strictly time-synchronized to the BS or when an external device is triggering the traffic arrival such as an industrial production line. Such impact of misalignment is illustrated in Figure 3-9. Upon misalignment, the resulting packet will be delayed until the next-coming slot, which may lead to damage or inefficient production in isochronous multi-robot production lines [D2.1]. In such cases, the tolerable latency is much shorter than the periodicity interval, but the trigger may still depend on some physical world trigger, meaning that the timing cannot be dictated by the base station.

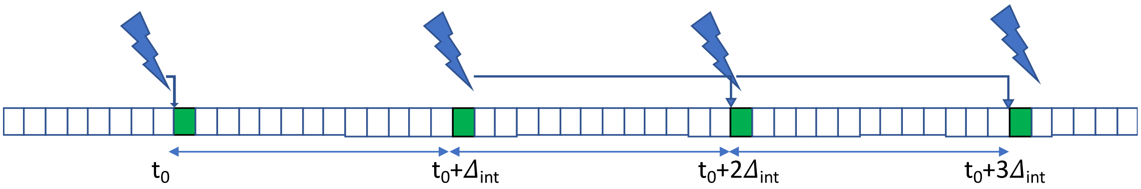


Figure 3-9 Misaligned arrivals cause increase in latency when an assigned grant (green) is missed.

The periodicity of assigned slots must be set based on the tolerable latency, so as to ensure that this is never exceeded. Consequently, the number of required slots per second for periodic assignment is simply $G_{per} = \frac{1}{t_{\text{lat}}}$, where t_{lat} is the tolerable latency. Since for short tolerable latencies the needed amount of slots can be quite high, we propose an adaptive scheme which is able to track the expected time of traffic arrivals. The proposed scheme assigns three slots around the expected arrival slot as illustrated in Figure 3-10. A preliminary first step is that the BS either needs to be informed about, or estimate the periodicity interval Δ_{int}. Thereafter, if an arrival is on time, i.e. if the corresponding packet is transmitted in slot 2, the BS keeps Δ_{int} as is. If, on the other hand, the arrival is early or late and is transmitted in slot 1 or 3, respectively, the periodicity interval Δ_{int} is adjusted accordingly. Thereby, minor variations in arrival periodicity, that do not exceed the duration of a slot per cycle, can be tracked by the BS. The required number of allocated slots per second for the proposed scheme based is: $G_{\text{track}} = \frac{3}{\Delta_{\text{int}}}$.

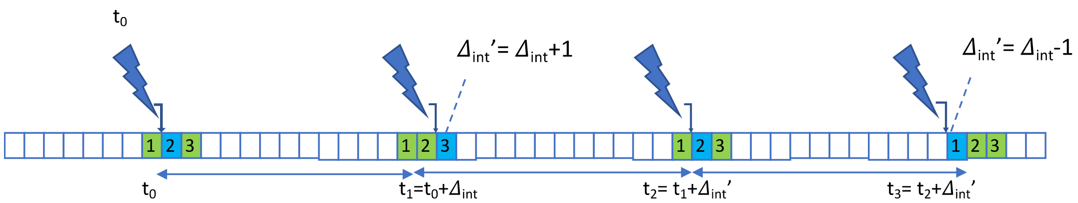


Figure 3-10 Proposed scheme for tracking traffic arrivals using configured grant slots.

Since the number of required slots of the periodic scheme depends on the tolerable latency and the number of required slots of the tracking scheme depends on the periodicity interval, we can compare the schemes in terms of the relation of these two parameters, as shown in Figure 3-11. For example, a traffic flow with arrival interval of 20 ms and latency limit of 1 ms, would require 20 slots per arrival using periodic assignment, or only 3 "tracking" grants. On the other hand, if the tolerable latency is more relaxed and larger than $\frac{\Delta_{\text{int}}}{3}$, then the periodic assignment is preferable from an overhead perspective.

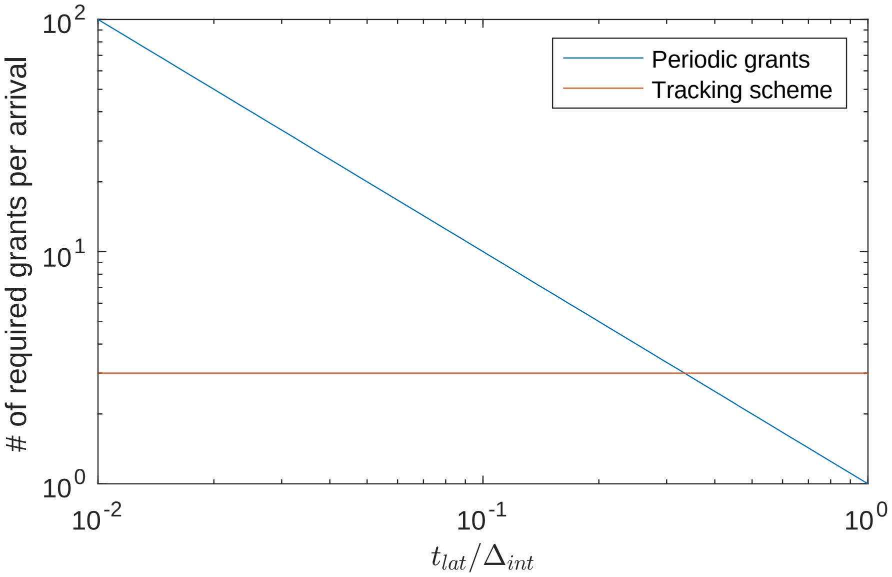


Figure 3-11 Comparison of required overhead in terms of ratio between tolerable latency and periodicity interval.

The proposed scheme has been evaluated numerically by means of Matlab simulations, where a realistic stochastic clock drift model has been used in a 5G NR URLLC configuration with TTI of 0.143 ms with representative traffic periodicities of tens and hundreds of milliseconds up to a few seconds. In all cases, the proposed scheme was found to eliminate the buffer waiting time, since a packet would always be transmitted in the following slot.

In summary, we have found that the proposed scheme is useful for reducing the number of needed slots compared to naïve periodic slot assignment. Specifically, for URLLC traffic, such as in the ONE5G core use case "*time-critical factory processes and logistics optimization*", with a latency requirement of 1 ms, the slot assignment overhead can be reduced to 1/3 or 1/30 for 10 ms and 100 ms cycle times, respectively, compared to the periodic slot assignment. This reduction in overhead allows the BS to support in the order of 3-30 additional uplink URLLC UEs, thereby improving service accessibility KQI.

Multiplexing of URLLC and eMBB service classes

Efficient solutions for multiplexing of diverse services such as URLLC and eMBB are obviously of importance. Efficient solutions that allow flexible and dynamic multiplexing on a shared channel help utilizing the available radio resources, subject to users QoS requirements, and therefore improve the maximum allowed offered traffic (i.e. less probability of blocking) and improve upon the KQI of service integrity. The following sections summarize developed solutions for multiplexing of URLLC and eMBB user traffic.

3.1.2.4 Fundamental scheduling policy considerations

In [KPM+19] we have presented fundamental considerations on efficient multiplexing of eMBB and URLLC. An efficient, yet simple and low complexity gNB scheduling policy has been developed that takes the following into account: (i) the users' radio channel conditions, (ii) priority and latency budget users, (iii) control channel overhead, (iv) HARQ, and (v) payload sizes to minimize effects of URLLC packet segmentation. We found that taking (i)-(v) into consideration is very important, yielding significant performance benefits over published scheduling algorithms in the open literature. The performance of the proposed scheduling algorithm is evaluated with an advanced 5G NR compliant system level simulator with a high degree of realism. Simulation results show promising gains of 98% latency improvement for URLLC traffic and 12% eMBB end-user throughput enhancement as compared to conventional proportional fair scheduling. For further details on this study, we refer to [KPM+19].

3.1.2.5 Multi-user MIMO null-space preemptive scheduling

In an earlier study (see ONE5G D3.1 [ONE18-D31] and [PPS18]), we developed the concept of preemptive scheduling for multiplexing of eMBB and URLLC type of services. Essentially this allows an ongoing downlink eMBB transmission to be overwritten by urgent URLLC transmissions, typically performed with a short TTI size (mini-slot) as compared to the eMBB transmission (slot-based transmission). Preemptive scheduling (with related recovery mechanisms [PPS18]) is an attractive solution for cases with moderate numbers of gNBs and UE antennas (say up to four antennas). However, for cases with more gNB and UE antennas, the earlier concept of preemptive scheduling can be further extended to more efficiently exploit the spatial dimension. That is, instead of completely overwriting an ongoing eMBB transmission when an urgent URLLC traffic arrives, one can attempt to instead multiplex those in the spatial domain.

Towards achieving this, we have first derived a basic multi-user multi-input multi-output (MU-MIMO) scheduling scheme, where a single eMBB user is scheduled per physical resource block. When URLLC traffic arrives, we attempt to co-schedule it with an existing eMBB transmission by using enhanced MU-MIMO pairing techniques. If no eMBB UE is found that can be paired with the URLLC UE, traditional preemptive scheduling is applied (i.e. overwriting part of the eMBB users' transmission as needed to serve the URLLC user). Here, the enhanced MU-MIMO pairing includes constraints to protect the URLLC users from intra-cell co-channel interference from the co-scheduled eMBB transmissions. This is achieved by utilizing subspace linear algebra techniques to control a user separation of co-scheduled eMBB and URLLC users, by having the targeted eMBB user(s) transmissions spatially projected over a pre-defined reference spatial subspace. This allows the URLLC users to align their data receptions to be within a possible null space of the reference eMBB spatial subspace, thereby, experiencing almost no inter-user interference from the eMBB transmission. As for the original preemptive scheduling concept, a mechanism to limit the impact on the victim eMBB users is provided, where the network informs the eMBB user(s) that they are co-scheduled with URLLC user(s), such that they can estimate their original transmission subspaces, leading to a significantly improved eMBB and ergodic capacity, while still preserving the URLLC QoS targets.

The associated signaling procedure of the developed MU-MIMO preemptive scheduling concept is summarized in Figure 3-12. First, the eMBB user(s) are scheduled with a TTI size of at least one slot. Once urgent URLLC traffic arrives, it is immediately co-scheduled on mini-slot resolution. When being co-scheduled with an eMBB user, the precoding of the eMBB user is altered to be in the predefined eMBB subspace, and the URLLC user is scheduled in the corresponding subspace. At the end of the eMBB transmission, the eMBB user is informed of the altered gNB precoding via a scheduling grant, such that it can take into account in the decoding process, and therefore partly mitigate the effect of suddenly being co-scheduled with a short URLLC transmission. For more details on the associated gNB-UE signaling procedures and the gNB algorithms, see [EP18c] [EP18d]. We name this concept *MU-MIMO null-space preemptive scheduling*.

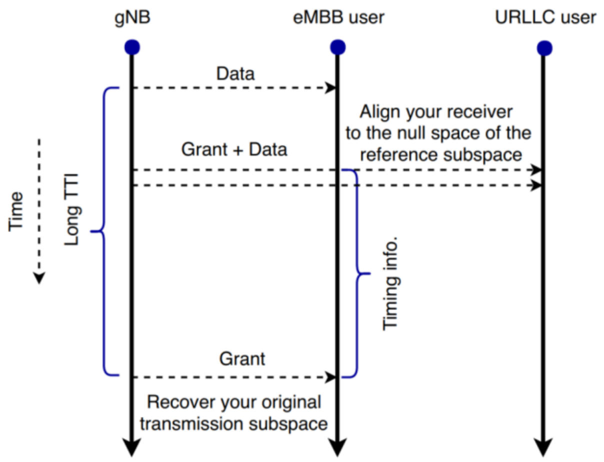

gNB
eMBB user
URLLC user
Data
Align your receiver to the null space of the reference subspace
Grant + Data
Time
Long TTI
Timing info.
Grant
Recover your original transmission subspace

Figure 3-12 Signaling diagram for the proposed MU-MIMO null space preemptive scheduler

The performance of MU-MIMO null-space preemptive scheduling has been extensively studied by means of 3GPP NR compliant dynamic system level simulations. Simulations have been conducted for a macro-cellular deployment with 500 meters inter-site-distance, including the full 3D urban macro propagation model, 8 antennas as the gNB per cell, and 2 UE antennas. Simulations include all the major RRM algorithms and the new NR physical- and MAC-layer design, etc. Different traffic loads have been considered, expressed as Ω(n,m), where n is the average number of eMBB users/cell and m denotes the number of URLLC users. In the presented results here, simple best effort eMBB traffic without strict QoS requirements is assumed (but in [EP18c] we also report results for CBR eMBB users). URLLC users experience bursty traffic, where URLLC payloads of 50 bytes are generated according to a homogeneous Poisson point process (aka FTP model 3 in 3GPP). The results in Figure 3-13 show the latency at the 99.999% reliability level and the average cell throughput for different traffic load conditions. Results are shown for the proposed MU-MIMO null-space preemptive scheduling scheme, as well as for traditional preemptive scheduling, standard MU-MIMO pairing cases, and standard time-domain weighted proportional fair (TD-WPF) – giving higher priority to URLLC over eMBB users. From the presented results, it is observed that the proposed scheduler solution performs best. The URLLC latency target of 1 ms is fulfilled, while the experienced latency is higher for the other reference schedulers. Figure 3-13 also pictures the average cell throughput, showing that also on this metric, the proposed scheduler solution offers the best performance. More results and details on the system level performance investigations can be found in [EP18a][EP18b][EP18c][EP18d].

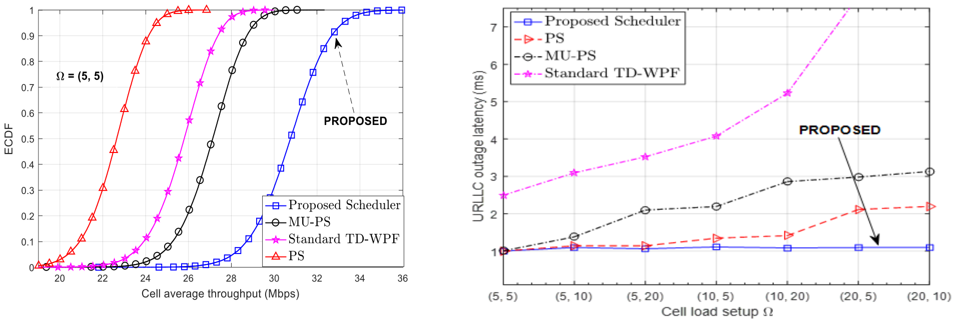


Figure 3-13 Null space MU scheduler: sample result of the URLLC outage latency and cell average capacity

In summary, we recommend the following:

- Use preemptive scheduling for efficient multiplexing of eMBB and URRLC user traffic with different TTI sizes for deployments with moderate numbers of gNB antennas in line with ONE5G D3.1, and now also supported in NR Rel-15 specifications.
- For deployments with more gNB antennas (say 8 antenna ports or more), the developed MU-MIMO null-space preemptive scheduling is recommended.
  - This is a promising method that exploits the spatial dimension to more efficiently multiplexing eMBB and URLLC users. It offers 60% capacity improvement.
  - For this solution to become a reality, it requires additional NR standardization of gNB-2-UE signaling to facilitate good isolation between co-scheduled eMBB and URLLC users (i.e. new DCI formats).
- The basic preemptive scheduling method (as now supported in NR Rel-15) and the newly developed MU-MIMO null-space preemptive scheduling scheme offer benefits in terms of improved KQI service integrity for both the URLLC and eMBB users, higher supported offered load per cell, so clear E2E performance benefits.

3.1.3 Centralized and multi-cell scheduling methods

We next present a set of promising scheduling schemes to fully unleash the performance potential of more advanced network architectures. We first present competitive centralized multi-cell scheduling algorithms for C-RAN cases with a lower layer split, i.e., the F2 interface as outlined in the earlier ONE5G deliverable, D3.1 [ONE18-D31]. Solutions for low complexity algorithms to enhance URLLC performance are presented – scheduling users from at most one cell per TTI – followed by more advanced CoMP based multi-cell cases for eMBB services.

3.1.3.1 Improving URLLC performance through centralized scheduling

Multi-cell scheduling algorithms for C-RAN architectures have, to the best of our knowledge, not been studied extensively for URLLC traffic cases earlier and are therefore addressed here. Our hypothesis is that C-RAN multi-cell scheduling solutions can significantly improve the overall performance, and thus contribute to enhancing the maximum offered URLLC traffic that can be tolerated, while still fulfilling the URLLC requirements. In this context, the most stringent requirement is the 1 ms one-way latency in the radio access network with a reliability of 99.999%. However, the 5G NR is also designed to support other classes of URLLC requirements as defined in the 5G QoS class indices (5QI) with latency budgets of, for instance 5, 10, and 20 ms, as well as reliability targets from 99% to 99.999% (see 3GPP specification 23.501). By being able to decide from TTI to TTI from which cell the different URLLC users are scheduled, we can reduce the probability of users experiencing queuing that risks jeopardizing their URLLC requirements. In line with 3GPP NR specifications, we assume a system that comprises the following [KPM+18]:

- UEs measure the RSRP from the cells, and perform CSI measurements on the Q strongest cells that are within a power-window of W dB (as compared to the strongest cell)
- CSI reports are subject to reporting delays before reaching the CU.
- The CU schedules UEs only from the cells that they report CSIs for, and not from multiple cells per time-instant (single point transmission). At each TTI, a UE is at most scheduled from one cell.
- Dynamic user scheduling, using a user-centric downlink control channel for transmitting the scheduling grant.
- Dynamic link adaptation with asynchronous HARQ with Chase combining is assumed.

In our search for C-RAN multi-cell URLLC scheduling algorithms, we prioritize solutions of modest complexity that are feasible for C-RAN architecture implementations, offering additional insight on the trade-offs between achievable performance and the use of sub-optimal algorithms with acceptable complexity. This is rather challenging given the considered system model, where the multi-cell scheduling problem can be formulated as an NP-hard linear integer optimization task. Solving this with a brute force algorithm in general has a complexity of $\mathcal{O}((C+1)^U)$, where U is the number of active users, and C is the number of cells. For the case where each UE can at most be scheduled from a set of Q candidate cells, the complexity reduces to $\mathcal{O}((Q+1)^U)$, which is still rather high complexity. Using the so-called matrix elimination method, it is possible to get the complexity down to $\mathcal{O}(U^3)$, and with our newly developed sequential algorithm of the matrix elimination solution [KPM+18], the complexity is reduced to approximately $\mathcal{O}(QU\log QU)$. The complexity of only $\mathcal{O}(QU\log QU)$ is rated as being reasonable, and implementable for the considered C-RAN architecture.

The developed scheduling policy also leverages the advantages and cost of applying packet segmentation of the small size URLLC packets. That is, without segmentation, only the full URLLC payloads of 50 bytes are scheduled, while for cases with segmentation, we allow that a URLLC payload is segmented to allow transmission over multiple TTIs. Cases without segmentation have the advantage of attempting to transmit the URLLC packets in one TTI, at the cost of not always being able to utilize all transmission resources as there may be insufficient resources to transmit full URLLC payloads. On the other hand, use of segmentation allows better utilization of radio resources, but at the cost of (i) higher PDCCH overhead as each transmission is accompanied with scheduling grant, and (ii) the possibility of errors at each transmission. The proposed C-RAN multi-cell scheduling solution therefore encompasses a component where at most one UE per TTI is subject to URLLC packet scheduling, making sure that all available PRBs can be utilized [KPM+18].

The proposed C-RAN multi-cell scheduling scheme has been evaluated in a state-of-the-art dynamic system-level simulator in line with the 3GPP NR Rel-15 specifications. The considered scenario is a traditional three-sector macro cellular network. Dynamic URLLC traffic is assumed, where URLLC payloads arrive according to a homogeneous Poisson point process. The URLLC packet size equals 50 Bytes, and the offered load is adjusted by varying the average packet arrival rate. The radio propagation model is according to the Urban Macro 3D channel model. Detailed modeling of radio resource management functionalities such as link adaption, Hybrid ARQ, scheduling (grant based), UE CSI reporting, etc. are included. The flexible NR frame structure is utilized by using short TTI transmission of 2-symbol mini-slots to ensure low latencies as per the URLLC use case.

Figure 3-14 (left) shows the complementary cumulative distribution (ccdf) of the URLLC packet latency for cases with traditional independent distributed (red curves) and the proposed centralized multi-cell scheduling methods (blue curves). Those results are achieved at a relatively high offered load. As can be seen, the results for the distributed cases experience a rather long tail of the packet delay. This is mainly contributed by gNB queuing delays when it happens that numerous URLLC packets need to be served from the same cell. However, when applying the proposed centralized scheme, the latency performance is significantly improved. This is achieved as the proposed multi-cell scheduling policy allows flexible scheduling of UEs from different cells on the fly, and thereby reduces the probability of experiencing queuing delays of the URLLC packets. In fact, it is observed from Figure 3-14 (left) that the tail of the URLLC latency distribution is reduced by approximately a factor of two at the low outage probability regime as is of interest for URLLC use cases, when compared at one high offered traffic load. Notice furthermore that the best performance is achieved when the scheduling is based on both throughput (TP) and packet delays (TP-Delay) as compared when using traditional max-TP scheduling. Figure 3-14 (right) compares how high offered traffic can be tolerated, while still fulfilling a certain URLLC quality target at a given delay target and reliability level. From this figure, it is observed that, by using the proposed centralized, multi-cell scheduling policy, the maximum tolerable offered URLLC traffic can be increased by approximately 30%-60%.

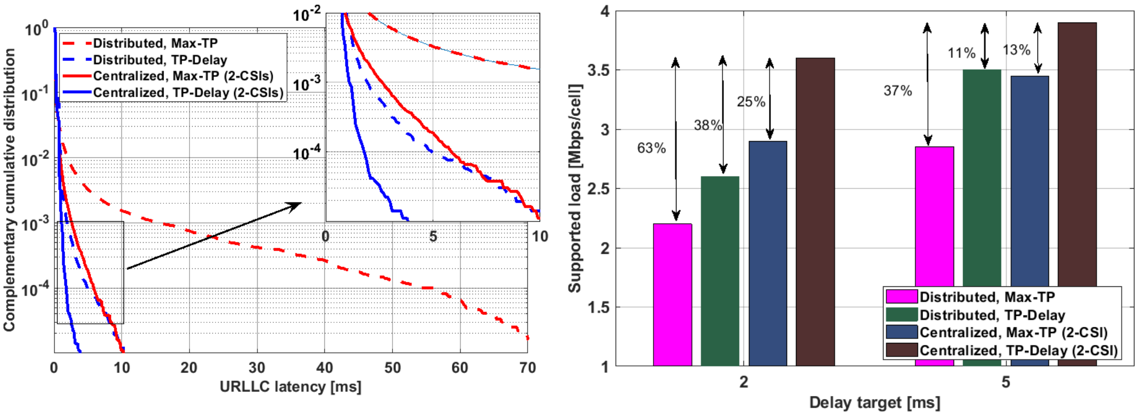


Figure 3-14 URLLC system-level performance for distributed and centralized network architectures. Left: ccdf of the URLLC packet latency, Right: maximum supported URLLC traffic load while still achieving a certain delay target. UEs at most report CSI from two cells.

In summary, our observations and recommendations are:

- Even a low complexity C-RAN multi-cell scheduling algorithm, where users are scheduled at most from one cell per TTI, offers significant URLLC performance improvements. The main benefit is the reduction in the probability of queuing delays.
- It is recommended to apply such solution as this helps improve on the KQI service integrity (and thus the E2E URLLC) performance, and also the maximum aggregated offered URLLC traffic that the system can tolerate without violating the latency and reliability targets (up to 60% gain).
- Those benefits are achieved for cases where UEs at most report CSI measurements from only 2-3 cells (low overhead) that fall within a received power window of 6 dB (measured on UE RSRP).
- Segmentation of URLLC packet is beneficial if limited to at most segmenting one UE per TTI to reduce the additional PDCCH scheduling overhead that comes with segmentation.

For more details, we refer to [KPM+18]. Moreover, also exploring multi-cell scheduling, the study has also been conducted by ONE5G, reported in [KPM19], where we have extended the concept of DPS to URLLC NR cases by developing new low complexity resource allocation algorithms. Extensive 5G NR system-level simulation results show that joint DPS and frequency-selective scheduling achieve 35% improvement of URLLC latency. This DPS scheduling variant is essentially a simplified version of the more extensive centralized multi-cell scheduling schemes that also takes load variations into account. We refer to [KPM19] for more details.

3.1.3.2 Advanced CoMP-alike centralized scheduling schemes for eMBB services

In the previous deliverable [ONE5G-D3.1], a detailed description of the centralized multi-cell scheduler proposed to handle eMBB traffic in Megacities scenarios was presented. Nevertheless, in this section we will recap some of the main concepts tackled by the centralized multi-cell scheduler.

The main objective of this scheduler is to efficiently perform scheduling decisions in a complex centralized radio access network (MAC-layer split or below), where several RRHs are connected to a single Central Unit (CU), with the aim of boosting overall system capacity. In order to do that, it is assumed that the CU will have full knowledge of channel conditions among all cross-links existing in the scenario, i.e. each pair UE-RRH, in such a way that it can use that information to allocate the users to the bandwidth parts where channel conditions are the best.

One of the intentions of the centralized multi-cell scheduler is to avoid penalizing the users located at cell-edge, even though their channel conditions are not the best. Thus, the CU will use a 3D-table populated with Proportional Fair (PF) metrics during scheduling decisions for the sake of fairness among the users.

Additionally, the scheduler implements advanced coordination techniques in order to manage sensitive situations where huge interferences are detected, for example, when a user is located at the intersection point of the coverage area of several cells. The coordination methods implemented are based on Joint Processing CoMP and NOMA techniques and they also allow frequency reuse by means of coordinating transmissions among the cells.

Hence, the objective of this work is to show the performance of the proposed algorithm by means of running simulations over a System-Level-Simulator (SLS), such as NS3, and compare the results to a distributed scheduler, following the same principle, in order to assess benefits, gains and cons of the proposed solution. The baseline distributed scheduler is shown in Annex D.

Several simulations considering different random users' locations have been launched in order to assess the performance achieved by this scheduler. Through applying Monte Carlo method is possible to quantify the real improvement by means of comparing the cell-aggregated throughput in both C-RAN and D-RAN scenarios over the same channel conditions.

As it can be seen in Figure 3-15 and Figure 3-16, the improvement in terms of cell aggregated throughput is greater in a C-RAN scenario compared to a D-RAN topology. .

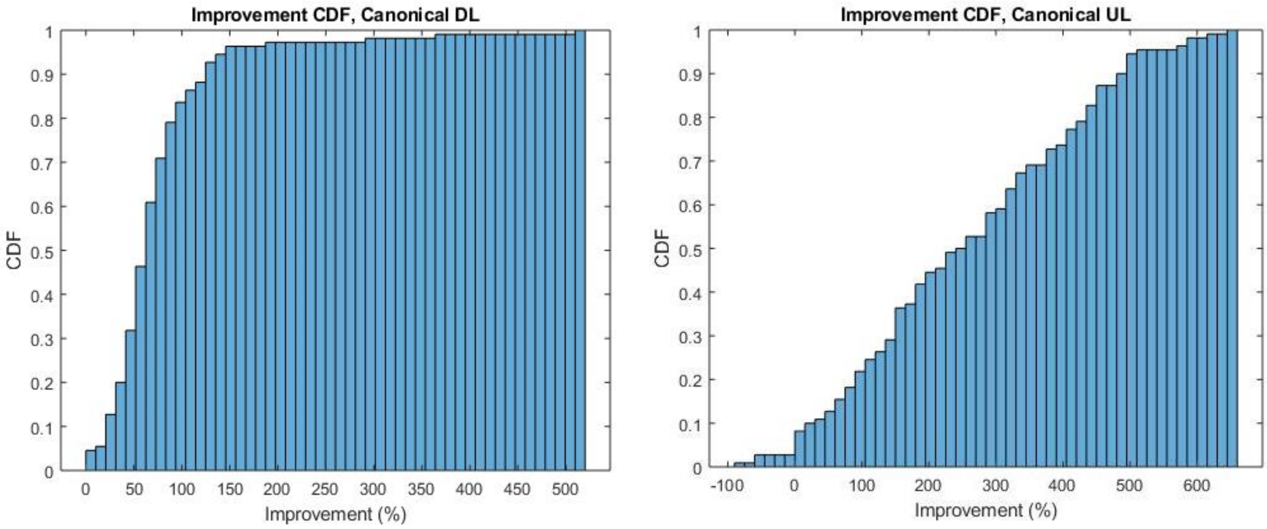


Figure 3-15 C-RAN Scheduler Vs. D-RAN (Canonical Scenario)

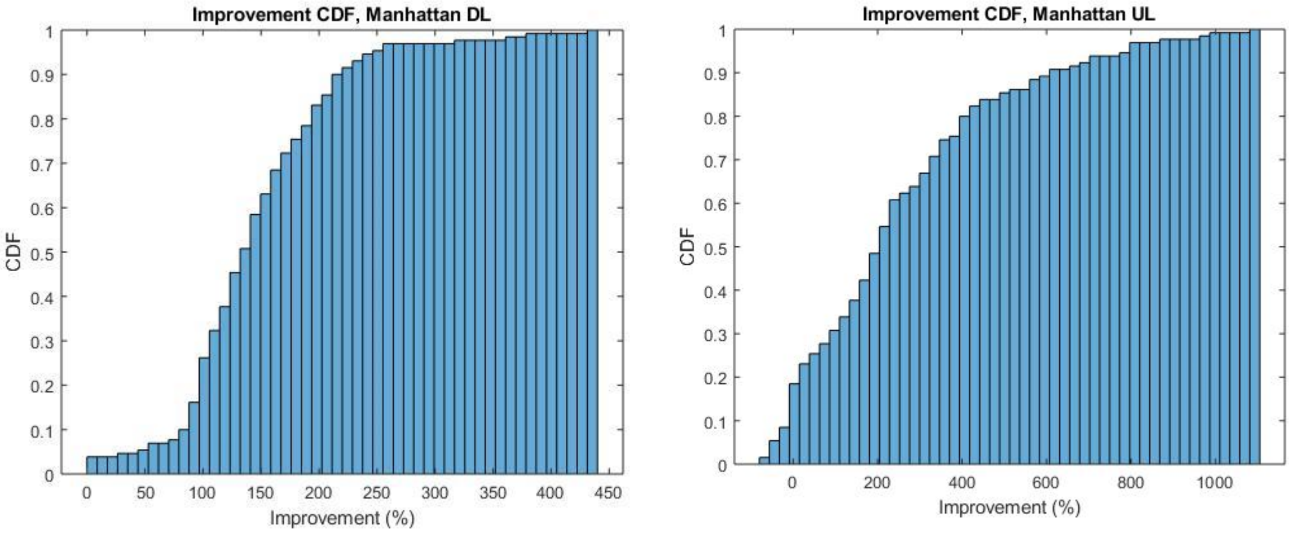


Figure 3-16 C-RAN Scheduler Vs. D-RAN (Manhattan Scenario)

The average improvement achieved by the use of a centralized architecture is shown in the Table 3-1.

Table 3-1 Average improvement for C-RAN vs D-RAN comparison

	Downlink	**Uplink**
Manhattan	147,12%	259,45%
Canonical	77,04%	261,75%

Looking at the Table 3-1 it can be noticed that a C-RAN scheduler introduces a significant improvement in the overall system capacity compared to a D-RAN scheduler. The resulting intelligent inherit in the C-RAN scheduler, by having in the CU all the information about all the channels (eNBs-UEs), allows to know the real interferences that would really happen in the scenario during the transmission time, and in turn, adjust the MCS (modulation coding scheme) to the SINR value that will be most likely seen at the receiver. In such a way, highest modulation orders would be used, in general, in the centralized scenario.

It is important to remark that this C-RAN solution is better in almost all situations, but it could be some situations, in which the algorithm does not provide such level of improvement, in terms of cell-aggregated throughput, as may happen in UL side:

The baseline UL D-RAN scheduler relies on ensuring full bandwidth occupancy, i.e. all the subbands are filled in each base station. But in the case of the UL C-RAN scheduler it cannot be ensured that each base station would fill all its resources because of the allowance of applying frequency reuse algorithms. Therefore, the improvement is correlated to the number of users that are located in each base station. As it can be seen in Figure 3-17 for the downlink, the dependency between the number of users and the improvement is almost flat, but in the uplink side, the improvement depends on the number of users, being lower if the number of users decreases in one base station, which could be even negative for low values, such as, for instance, 3-6 users.

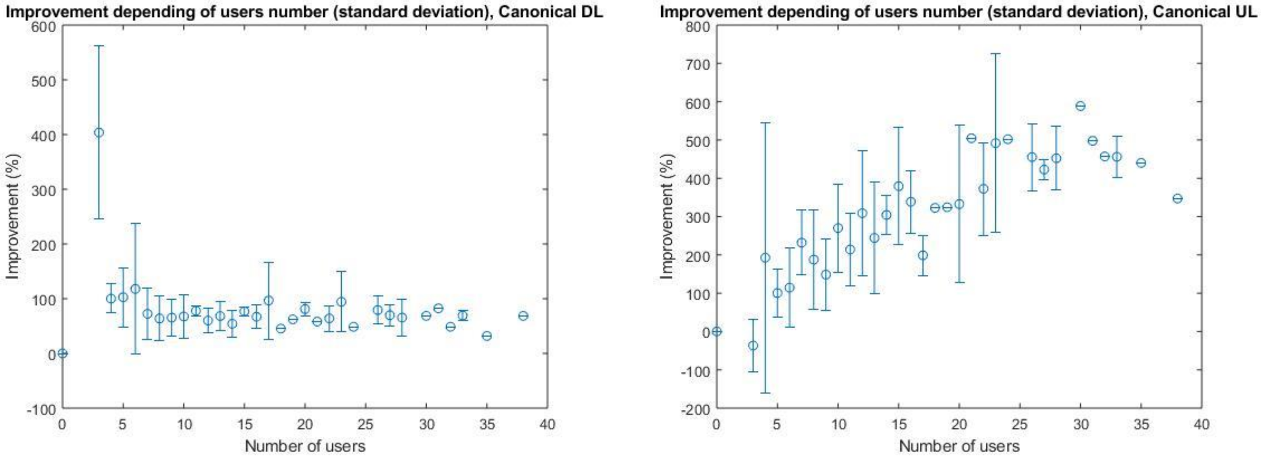


Figure 3-17 Improvement dependency with the number of users

One of the main challenges of a scheduler is to ensure that all the users have the chance to obtain radio resources, even if they suffer from very bad channel conditions. For this purpose, the C-RAN scheduler presented in this section is based on PF metrics, which depends not only on the achievable throughput at the current TTI but also on the historical throughput for each UE. By the use of PF metrics, it can be ensured that all the users would obtain resource allocation during the simulation time.

As it can be seen in the following figure, all the users obtain resources from the scheduler. The graphs represented at the top of the figure show the percentage of RB (resource blocks) allocations that has been granted to each UE (imsi). The lower part of the figure illustrates the same concept but ordered in an ascendant way. It can be noticed from both graphs that users without transmissions do not exist.

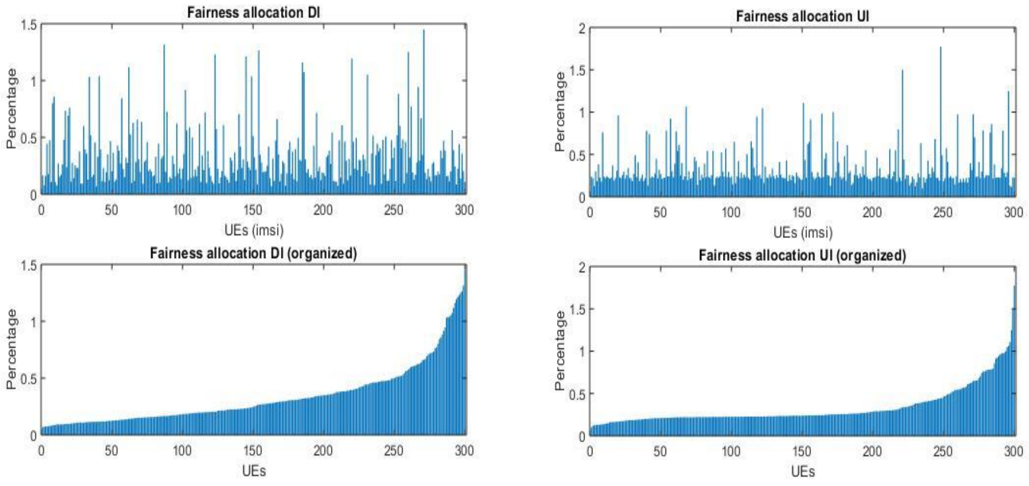


Figure 3-18 Fairness obtained with C-RAN scheduler

3.2 Multi-channel access solutions

In this section we will further investigate the most promising Multi-Channel Access (MCA) solutions for the 5G New Radio (NR) multi-service scenario, namely dual/multi-connectivity (DC/MC) and carrier aggregation (CA). In general, MCA techniques can be classified based on the following key characteristics that will influence their performance and complexity [MLP+18]:

- Which radio protocol layer is hosted by the anchor point that schedules the data, and whether resource allocation and transmission coordination can be leveraged;
- Whether the resource aggregation happens within the same node or across distinct nodes, implying in the latter case the availability of a fast network interface to exchange information among the nodes;
- At which frequency bands the radio aggregation occurs: intra-frequency vs. inter-frequency schemes.

As per 3GPP, DC/MC is an example of Packet Data Convergence Protocol (PDCP) based data handling (splitting or duplication) and independent usage of the resources at distinct nodes. The operations and enhancements for the efficiency of MC with packet duplication will be investigated in Section 3.2.1, whereas solutions for secondary cell selection with MC will be proposed in Section 3.2.2.

3.2.1 Multi-connectivity with packet duplication: operations and enhancements

5G NR URLLC support defined in Release 15 comprises a set of features to ensure the stringent reliability and latency targets. PDCP-level packet duplication is one such important feature. Duplicating the data through the same cell group (CG) as in CA, or different CGs as in DC/MC, allows the reception of multiple copies of the same data, thereby, improving the reliability through diversity and repetition. The basic illustration of the feature in the context of NR dual connectivity (NR-DC) is shown for the downlink case in Figure 3-19.

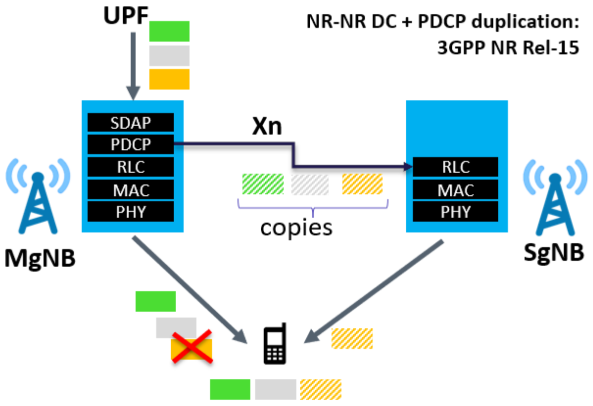


Figure 3-19 Schematic of Dual Connectivity with PDCP Duplication for downlink transmissions, with a user-plane function (UPF) sending data to a terminal via a Master 5G-NR gNB (MgNB) and a Secondary 5G-NR gNB (SgNB).

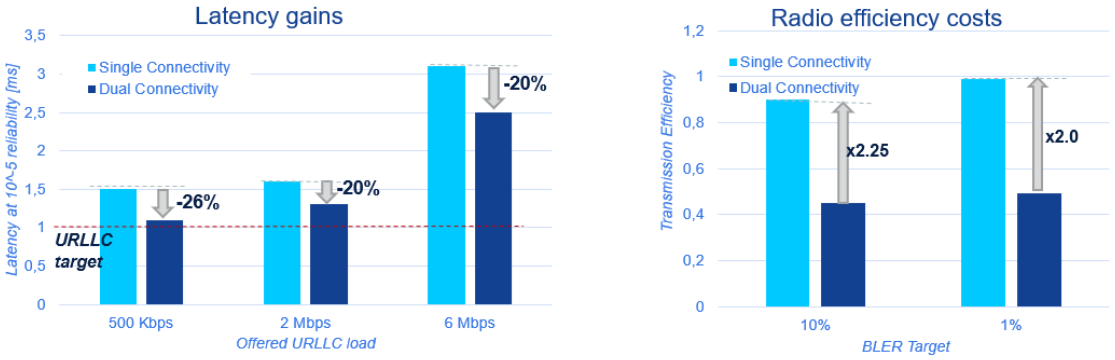
However, the improved performance is obtained at the expense of an increased number of transmissions in the network, and, consequently, an increase in cell load, interference level and queueing delays. Furthermore, the additional resources used for duplication are unnecessary most of the times, i.e. when the primary transmissions are successful. As an example, if the BLER target for first transmissions is set to 1%, on average, 99% of the duplicate transmissions will be redundant. Table 3-2 summarizes the trade-off between gains and cost associated to data duplication.

Table 3-2 Summary of gains and cost associated to data duplication

Data duplication gains for a given UE	**Data duplication costs for the network**
⇒ Improved latency and reliability:	⇒ Increased radio resource consumption

Under independent transmission from both nodes $P_{out,DC} = P_{out,1}P_{out,2}$ where $P_{out,i}$ is the outage probability through node i.	⇒ Increased interference ⇒ Increased buffering delay

The baseline system level performance achieved with Release 15 URLLC when adopting packet duplication is depicted in Figure 3-20. This performance result was obtained via simulations conforming to the assumptions given in Annex B of [ONE18-D31], (i.e. using a 3GPP heterogeneous network (HetNet) scenario with a micro underlay and a macro overlay layer). We remark that the 1-ms latency target can be achieved only for low offered URLLC load because the considered scenario is rather challenging due to the background eMBB traffic creating full interference. It can be seen that packet duplication can achieve a significant improvement in the service integrity KQI (KQI-sI) of URLLC applications, i.e. user-plane latency at the 5-nine reliability (i.e. 99.999%): the gain in latency reduction is about 20% regardless the URLLC offered load. However, this comes at the cost of roughly doubling the radio resources consumed, confirming that enhancements to the radio resource efficiency of packet duplication should be introduced to better reap the gains and avoid wasting large amount of radio resources, which can potentially outweigh the benefits. Further simulation results are presented in [MLL+18]. In addition, a theoretical analysis of the outage probability gains with MC is presented in Annex B (Section 7.2). The obtained analytical results indicate that DC/MC results in significant outage probability reduction compared to single connectivity schemes. The findings can be used to determine the BLER targets required to achieve the URLLC outage probability target. For example, the desired URLLC reliability can be achieved with more than 10 times higher BLER target (for individual transmissions) for DC, as compared to SC schemes.



URLLC latency in 3GPP HetNet scenario 2A [TR 36.872]
System-level simulations, ***URLLC + background eMBB traffic***

Transmission efficiency for single-UE scenario
(I.e. avg # of received PDCP PDUs per PHY transmission)

Figure 3-20 Illustration of the URLLC latency gain and transmission efficiency of packet duplication

Proposed enhancements for resource efficiency

To tackle the challenges mentioned above, in this study we will identify various mechanisms to increase the radio-resource efficiency of packet duplication.

For instance, an effective means to avoid unnecessary duplicate transmissions (i.e., transmissions of packets which were already received successfully at UE) is to ensure faster discarding of unnecessary copies which are buffered at the network side. In fact, in Release 15 NR, signaling of discard indication and successful delivery indication are supported over Xn/F1 network interfaces, based on the reception of the UE HARQ acknowledgment by the corresponding transmitting node [3GPP38.425]. In principle, this makes PDCP protocol data unit (PDU) flushing at lower layers possible; however, such indications may be available too slow in light of the tight delay budget of URLLC (i.e. 1 ms). Depending on the deployment, the redundant packets may have already left the buffer and been transmitted when accounting for realistic constraints such as processing and signaling latency delays. Therefore, as a first enhancement, we propose a network discard mechanism that relies on a novel UE duplication status report [NLL+18], [3GPPR2-1817582]. This report indicates in a timelier manner to other node(s) in the duplication set that a certain PDCP packet has been successfully received, thus allowing the duplicating nodes to discard the flagged PDCP packet if it has not yet been transmitted. This report makes in-network discard feasible under realistic constraints.

In addition, the presence and prioritization of duplicate packets may result in queuing delay for other traffic and/or users depending on the load conditions. This may be undesired from the perspective of achieving optimal overall system performance. According to Release-15, the duplicating node is aware of whether a certain PDCP PDU is a duplicate or not and could apply some differentiation based on this information. However, it may be beneficial that the node hosting PDCP provides assistance information to the node hosting the secondary RLC entity to adjust scheduling decisions and radio resource allocation. Therefore, according to the second enhancement we propose assistance information in terms of e.g. BLER target to apply to the transmissions of duplicates at the SgNB, with the aim to relax the amount of resources spent for duplicate transmissions and limit the impact on other traffic [3GPPR2-1817582].

Similarly, as third enhancement a novel signaling framework is proposed to convey additional information between the network entities to significantly reduce the total amount of packet duplicates transmitted to the UE. Particularly the framework should allow means to indicate to hold back a transferred duplicate packet at a node until a further indication is received. At the reception of this indication the corresponding node should either timely transmit or discard the duplicate packet [3GPPR3-186693]. For instance, this could bring advantage if as soon as a transmission from the primary node failed (i.e. the HARQ NACK is received for a PHY transmission associated to a PDCP PDU), the secondary node(s) transmitted the same PDCP PDU (in addition to the primary node’s HARQ retransmission) to increase the reliability of the failing packet. This scheme would drastically reduce the number of duplicates because, as discussed above, the transmission failure in URLLC scenarios will be typically very limited (i.e. 1% of first transmissions will fail when the actual BLER is 1%).

In the following of this section, we will evaluate the gain when adopting a selection among the proposed enhancements, namely i) fast duplicate discarding and iii) selective duplication. It should also be noted that these solutions are in scope of the 3GPP study RP-182090, Study on NR Industrial Internet of Things (IoT), [3GPPRP-182090]. The study comprises, among its objectives, L2/L3 enhancements towards data duplication and multi-connectivity enhancements. It aims at investigating improved resource efficient PDCP duplication (e.g. through coordination between the nodes for PDCP duplication activation and resource efficiency insurance, by avoiding unnecessary duplicate transmissions etc.). Besides, it covers PDCP duplication with more than two copies leveraging (combination of) DC and CA, whereupon data transmission takes places from at most two nodes. In [3GPPR3-185547], initial solutions along the proposals discussed above were also contributed to 3GPP RAN Working Groups.

System-level performance evaluation of PDU discard indication

Figure 3-21 shows the proposed mechanism for fast duplicate discarding: in this example, as soon as the UE correctly receives the PDCP PDU from the Master Node, it sends a discard indication to the Secondary Node in order to prevent transmissions of redundant duplicates.

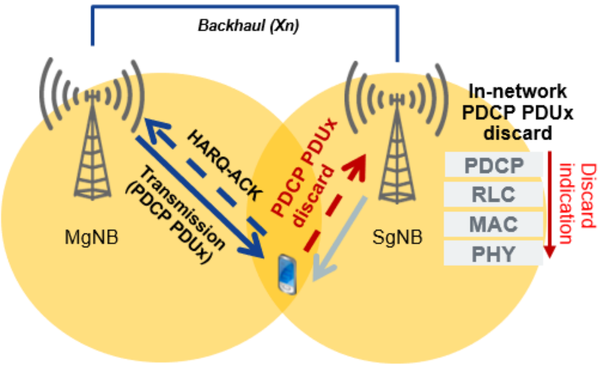


Figure 3-21 Illustration of the proposed PDU discard indication method

The performance of the proposed approach has been evaluated by means of system-level simulations conforming to the assumptions given in Annex B of [ONE18-D31], except for disabling the pico layer. This results in an intra-frequency deployment, where the duplicates transmitted either by adjacent sectors of the same gNB or by different gNBs towards a UE may interfere with each other, implying that duplication must be carefully used. The simulation results are provided in Figure 3-22. In the simulation without eMBB background traffic and 1-Mbps URLLC offered load (left-hand side subfigure), we notice a latency gain of 20% thanks to the proposed discarding mechanism. On the other hand, in the simulation with eMBB background traffic (right-hand side subfigure), we obtained a latency gain of 12% despite increasing the offered URLLC load from 1 Mbps to 2 Mbps.

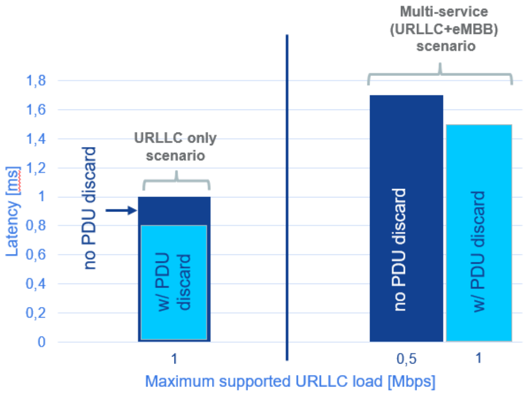


Figure 3-22 Simulation results of PDCP duplication with and without enabling PDU discard indication from the UE

System-level performance evaluation of selective duplication

Figure 3-23 shows the proposed mechanism for selective duplication: in this example, as soon as the Master Node receives a HARQ NACK indication from the UE, it sends a “urgent scheduling request” indication to the Secondary Node, triggering it to send its duplicate of the PDCP PDU, which previously was on hold.

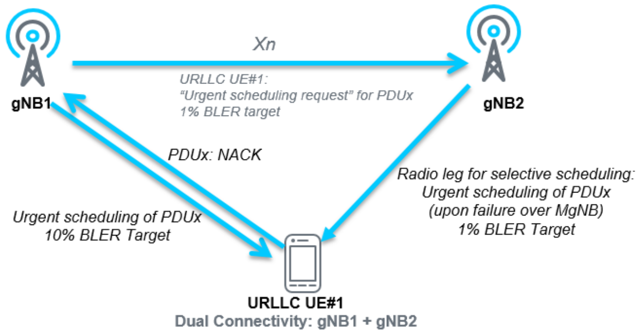

Xn
gNB1
gNB2
URLLC UE#1:
"Urgent scheduling request" for PDUx
1% BLER target
PDUx: NACK
Urgent scheduling of PDUx
10% BLER Target
Radio leg for selective scheduling:
Urgent scheduling of PDUx
(upon failure over MgNB)
1% BLER Target
URLLC UE#1
Dual Connectivity: gNB1 + gNB2

Figure 3-23 Illustration of the proposed selective duplication approach

The performance of the proposed approach has been evaluated by means of system-level simulations conforming to the assumptions given in Annex B of [ONE18-D31] (inter-frequency HetNet scenario), assuming a variable offered load for URLLC traffic and full-buffer eMBB background traffic. In particular, a performance comparison of single connectivity, PDCP duplication without the selective duplication feature (referred to as "blind duplication"), and PDCP duplication enabling the selective duplication feature (referred to as "selective duplication") for a fixed URLLC offered load of 3.5 Mbps is provided in Figure 3-24. The blind duplication results are obtained enabling the fast duplicate discarding feature discussed previously. From the figure, it can be seen that selective duplication is the only approach that clearly achieves a reliability of 99.999% within the 1-ms URLLC delay budget (see left-hand side subfigure). The target URLLC performance is met by the proposed solution thanks to a massive reduction of consumed radio resources (see right-hand side subfigure), with the selective duplication approach matching the number of RBs consumed by the single connectivity in the clear majority of URLLC packet transmissions. Indeed, we recall that selective duplication reduces the radio-resource consumption of a factor (1 – BLER) = 99% with respect to blind duplication, assuming BLER = 1%.

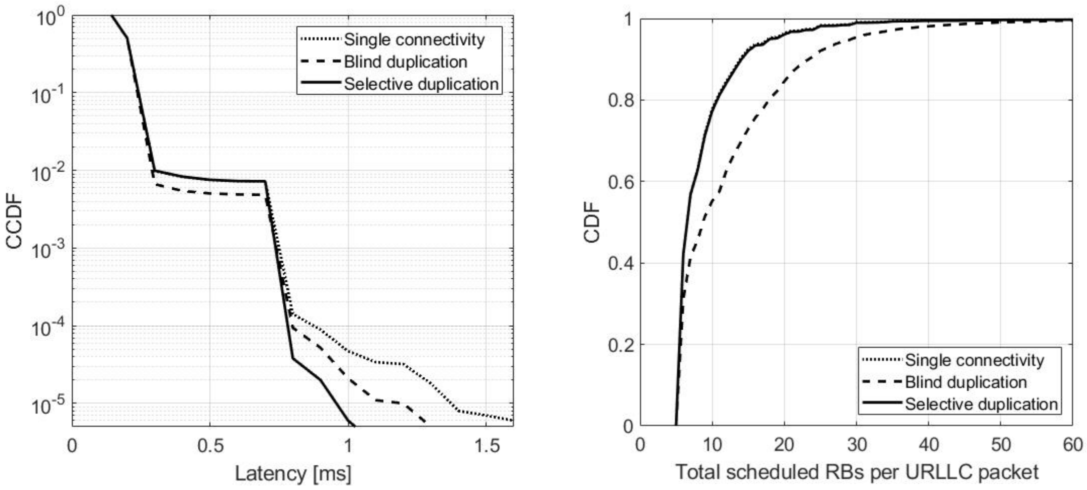


Figure 3-24 Simulation results of single connectivity and PDCP duplication, enabling ("selective duplication" label in the legend) and not enabling ("blind duplication" label in the legend) the selective duplication feature. A URLLC offered load of 3.5 Mbps and full-buffer eMBB background traffic are assumed for all approaches.

On the other hand, a performance comparison of single connectivity, blind duplication, and selective duplication for different offered URLLC loads and full-buffer eMBB background traffic is provided in Figure 3-25. It can be seen that the proposed approach can almost support 4 Mbps of offered URLLC traffic, outperforming the performance of single connectivity and blind duplication under 1 Mbps offered URLLC load. Thus, selective duplication increases the supported URLLC load of roughly 4 times with respect to single connectivity and blind duplication. The enhanced spectral efficiency obtained by the selective duplication of URLLC packets results also in an increased eMBB throughput especially for the macro layer, which achieves approximately the same performance as in single mode (benchmark).

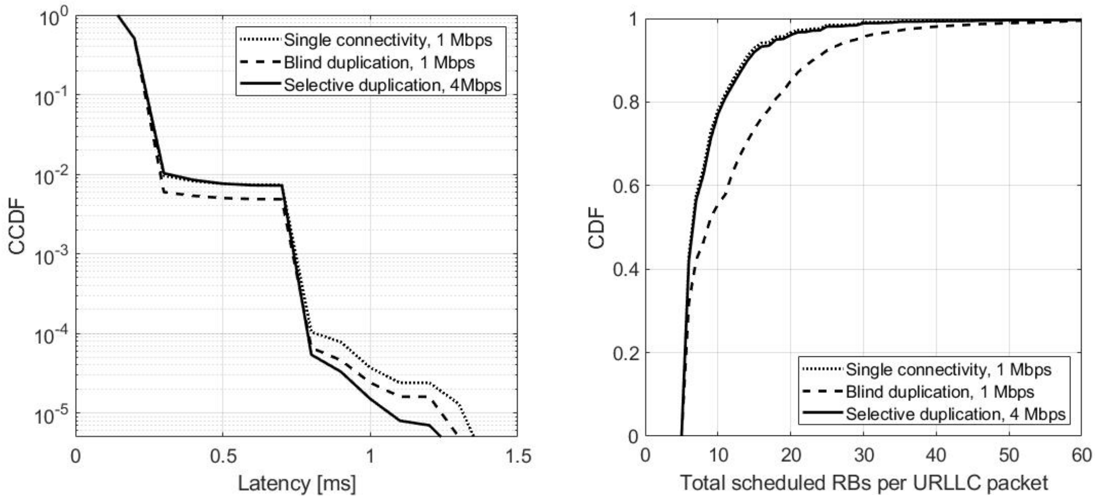


Figure 3-25 Simulation results of single connectivity and PDCP duplication, enabling ("selective duplication" label in the legend) or not enabling ("blind duplication" label in the legend) the selective duplication feature. A URLLC offered load of 1 Mbps and full-buffer eMBB background traffic are assumed for single connectivity and blind duplication, whereas a URLLC offered load of 4 Mbps and full-buffer eMBB background traffic are assumed for selective duplication.

On using more than two simultaneous copies and learning from higher layer duplication

A dedicated aspect under consideration in this ONE5G study is whether PDCP duplication supporting more than two copies at a time brings additional gain on top of using at most two copies at a time as per NR Release-15. The scenarios in scope assume maximum two nodes (according to the scope of the 3GPP study) and allow gNB_1 and gNB_2 having N_1 and N_2 component carriers (CC) respectively. This means that up to $N_1 + N_2$ copies of a PDCP PDU can be transmitted over gNB_1 and gNB_2.

Figure 3-26 shows the latency resulting from LTE drive tests {A, B, C, D}, which correspond to multiple UEs in the same device being connected to up to 4 network operators (see [LKP+18]). The packets are combined at the application layer comprising the cases of 2x, 3x, and 4x network combinations. Despite the difference in employing higher layer duplication (rather than PDCP duplication) and different nodes, the gain mechanisms are similar. Therefore, the learning from these results, as highlighted in the following, will be applicable to the scenario in scope (i.e. PDCP-level duplication with more than two copies at the time).

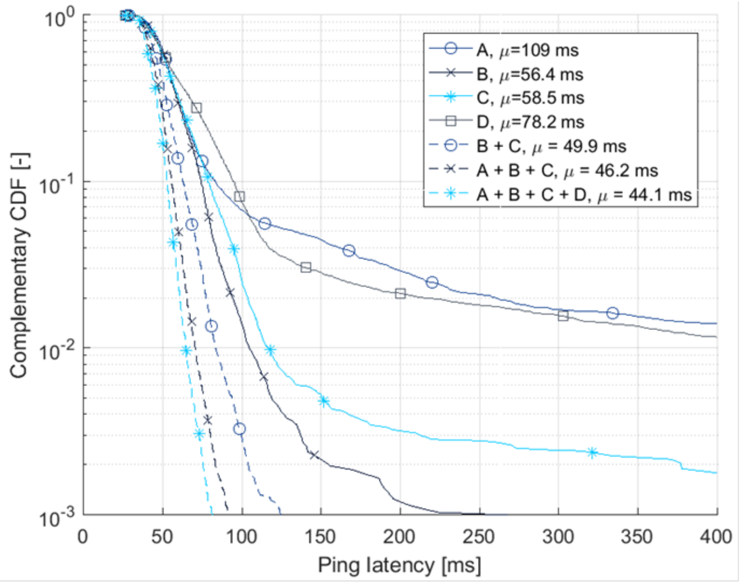
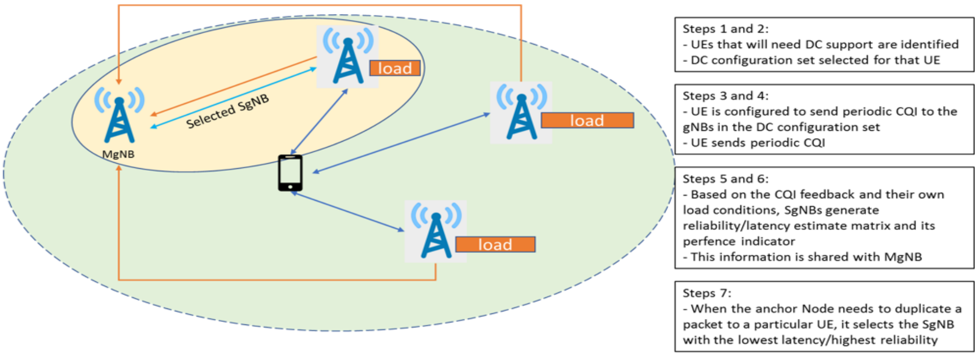


Figure 3-26 Illustration of latency gains assuming single network, 2x, 3x, and 4x networks, where μ denotes the average latency value (see [LKP+18])

Based on Figure 3-26, it can be observed that the case of 3x (and 4x) (i.e. which entails 3/4 copies at a time) provides benefit in the latency tails on top of 2x (i.e. which entails two copies at a time from the best two nodes). This is because the transmissions occur from 3 (4) distinct nodes, and therefore are entirely uncorrelated (i.e. adding further diversity as compared to the 2x case). Also, the test was done in a “single-UE scenario” where the costs of duplication to the other traffic are not visible. When restricting the scenario to having maximum 2 nodes involved in the duplication of more than two copies, the additional transmissions (on top of 2 copies) will have a large degree of correlation in terms of signal (i.e. similar slow and fast fading for same node in the neighbour CCs) and will increase the cost of duplication. Yet, the load level may differ across CCs, and these variations could be exploited by fast switching of the two RLC entities (“legs”) to use at a time. Therefore, it is anticipated that most of the diversity gain could be ripped off by allowing fast switching of maximum 2 copies at a time from the set of configured legs associated at the two nodes.

To this end, we propose a method to enable estimating and sharing with the primary gNB of the achievable reliability and latency of a transmission to a UE through an associated gNB. Such information can guide the PDCP entity in the primary gNB to select the most appropriate duplication set for a given UE at any given time, as shown in Figure 3-27. Basically, the PDCP entity would be able to determine on a fast basis the best 2 legs which can enhance the performance gains of duplication. The estimate is based on the UE reporting of multi-cell CQI/CSI to the primary gNB, i.e. reporting the quality both of the primary node/serving cell and secondary nodes/neighbour nodes.



Steps 1 and 2:
- UEs that will need DC support are identified
- DC configuration set selected for that UE

Steps 3 and 4:
- UE is configured to send periodic CQI to the gNBs in the DC configuration set
- UE sends periodic CQI

Steps 5 and 6:
- Based on the CQI feedback and their own load conditions, SgNBs generate reliability/latency estimate matrix and its perfence indicator
- This information is shared with MgNB

Steps 7:
- When the anchor Node needs to duplicate a packet to a particular UE, it selects the SgNB with the lowest latency/highest reliability

Figure 3-27 Illustration of latency signaling for fast “leg” selection

3.2.2 Automatic secondary cell selection for multi-connectivity

A solution that addresses the automatic allocation of CCs to a UE, encompassing both single-node and multi-node Multi-Channel Access (MCA), and relies on a rule-based system, is proposed below. This solution was previously presented (but not assessed) in Section 4.1.3 in D3.1 [ONE18-D31]. Although it is described in the light of maximizing the throughput of eMBB services, for which the data flow is split among the assigned CCs, it could be applicable to the URLLC case as well, where the data flow would be duplicated for reliability purposes.

Focusing on the solution, the proposed algorithm aims at determining the number and indices of CCs to be assigned to a specific UE, as well as the gNB(s) providing them, according to specific optimization objective(s) (e.g., maximize throughput, load balancing among CCs, etc.). As described in Section 4.1.3 in D3.1 [ONE18-D31], the proposed solution has been implemented using a fuzzy logic approach based on a set of IF-THEN rules. This technique is especially suitable to be applied to cellular networks since it allows operators to map their policies to the management functions. Using fuzzy logic terminology, the antecedents of the rules (the condition to be met in the IF part of a rule) are constructed based on performance information, gathered from the UEs (e.g., RSRP or the reference signal received quality, RSRQ) and the CCs themselves (e.g., load information). The consequents of the rules (the result of a rule when this is proved as true, that is, the term after the THEN part) are scores (standing for their suitability given a certain policy), over which an aggregation method is applied. Finally, the CCs with the highest aggregated scores, are assigned to the users, whenever they are above a minimum threshold.

A simulation study has been carried out in an environment of load imbalance. That is, a situation in which the heterogeneous spatial distribution of the users throughout the deployment area makes a reduced group of nodes support most of the offered traffic, leading to a high number of blocked connections, whereas many other nodes remain almost unused. To that end, the UE-reported RSRQ and the load level of a CC, derived from the instantaneous number of UEs allocated to a CC, have been used as input performance metrics. Different rules have been defined, assigning low scores to low values of RSRQ and high-load levels, and high scores in the opposite case. The final score of a CC is computed as the average of the scores provided by each rule.

The proposed solution has been tested using a system-level simulator in a macro-cell scenario, made up of 12 tri-sectorial sites with an irregular space distribution, where, in its turn, each sector is composed of five 1.4 MHz co-located CCs. A geographical region with a high density of users, spread throughout several sites, has been simulated to produce the load-imbalanced scenario. Two different sets of input parameters have been simulated: first, as a baseline, a case in which the rules only consider the RSRP as input, following the traditional approach for UE-gNB association, and second, the case in which both RSRQ and CC load metrics are assessed.

Figure 3-28 shows the 5th, 50th and 95th percentiles of the UE throughput for the baseline case with dashed lines and the ones of the proposed solution with solid lines. The achieved throughputs are shown as a function of the number of assigned CCs (one to five). A UE may be assigned all or some of the available CCs, depending on how many of them have a score above a pre-defined threshold. The proposed CC selection method results in a UE throughput improvement, as a result of efficient load balancing. More specifically, the proposed solution provides 100% average gain for the users experiencing the worst throughput values (5th percentile) over the state-of-the-art RSRP-based solution, and up to 75% gain at the peak throughput (95th percentile). As expected, it is also observed that higher numbers of CCs generally imply higher UE throughputs. In cases where more than one CC is assigned to a UE, each CC may be provided by a different node. For example, for the case of two CCs, approximately half of the UEs used CCs belonging to the same node (thus, applying CA), whereas the other half used CCs provided by different nodes (thus, applying DC).

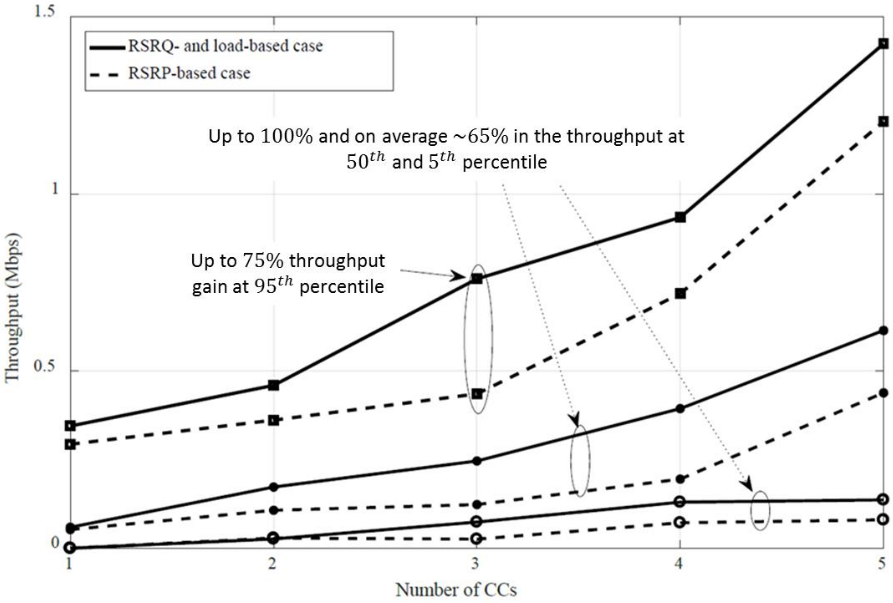


Figure 3-28 Per-UE achieved throughput for an increasing number of CCs; comparing the proposed solution for CC assignment with a traditional RSRP-based UE-gNB association

In order to complete the study, several types of eMBB services have been considered. In particular, FTP services (FTP), web browsing (WEB), video streaming (NRT) and real time video (RT). To analyse the performance of these services, QoE metrics are considered [OTL18][NLS10]. In addition, these QoE metrics have been included as input of the proposed CC manager algorithm. Specifically, when the CC manager is executed for a certain user with a certain service and for a specific CC, QoE metrics of other users with the same service that are allocated to that CC are considered.

Figure 3-29 compares throughput results for baseline (CCMRSRP), first version of the CC manager (CCMRSRQ) and the CC manager based on QoE metrics (CCMQoE), which is the one that allows to optimize throughput for all cases. Finally, Figure 3-30 shows the gain obtained for CCMRSRQ and CCMQoE compare to baseline. The metrics used in this case in the Mean Opinion Score (MOS), whose value represents the subjective service quality that a user perceives, for each service. It is important to point out that the technique used for the CC manager implementation (i.e., Fuzzy Logic) does not implies a significant increase of system complexity.

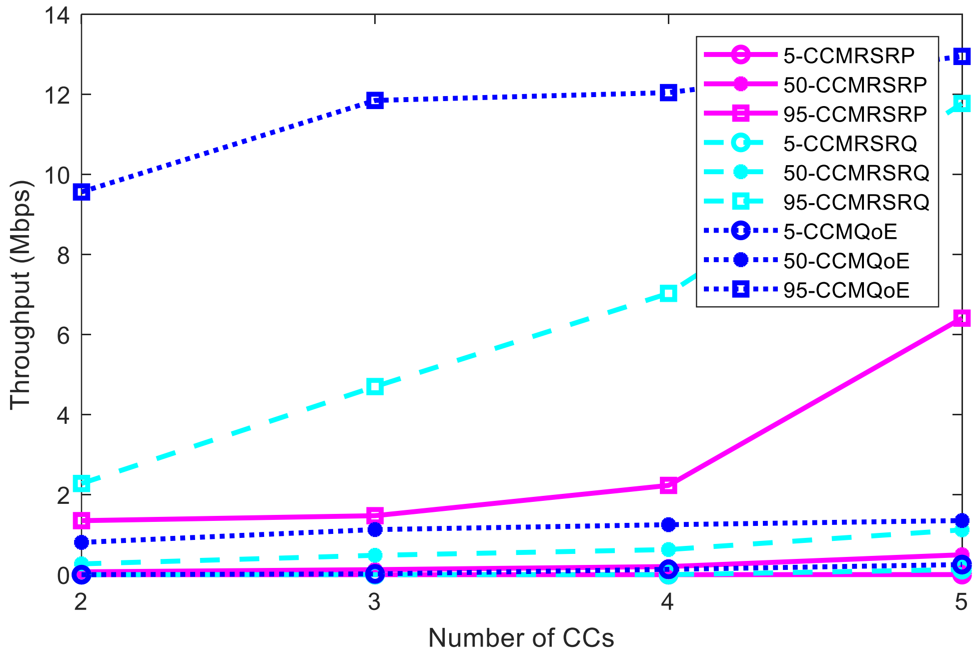


Figure 3-29 Per-UE achieved throughput for an increasing number of CCs; comparing both CC manager solution, one based on RSRQ and other one based on QoE, with baseline

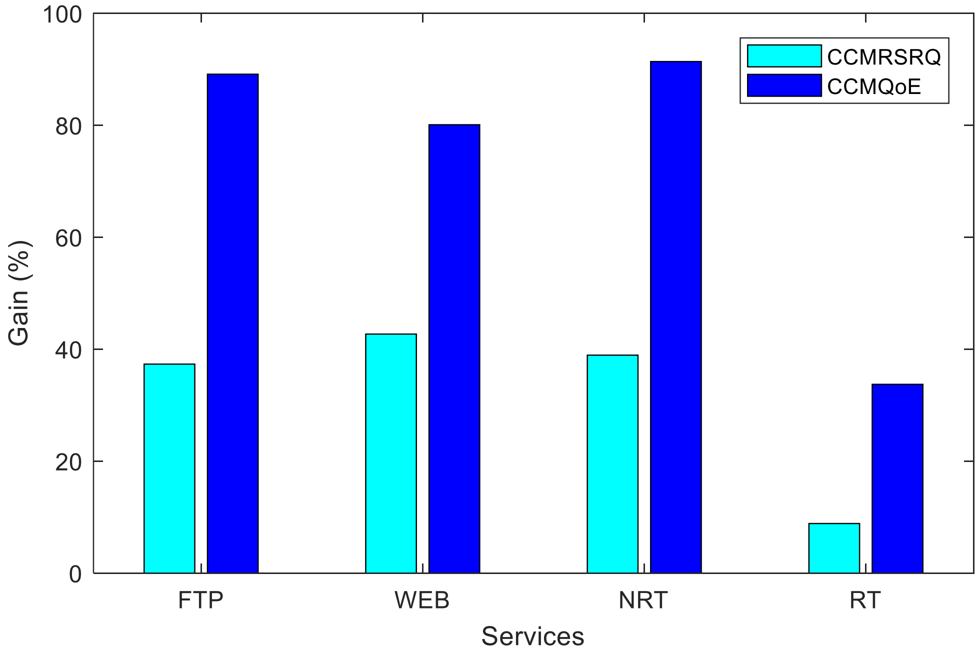


Figure 3-30 Gain of MOS for different eMBB services

Once the CC assignment is carried out, the traffic flow that is served by each CC needs to be determined. With this objective, a technique for traffic distribution among the CCs of the nodes that serve a MC-enabled UE is proposed. The target of the proposed solution is to find an enhanced traffic distribution with small signalling overhead that outperforms a conventional homogeneous traffic distribution.

Specifically, the amount of traffic carried by each CC is computed as the weighted sum of previous metrics (i.e., number of PRBs and RSRQ) as seen in the following equation.

$$T_i = W_{Load} \cdot x_{Load_i} + W_{RSRQ} \cdot x_{RSRQ_i},$$

such that

$$\sum_{i=1}^{N_{CC}} T_i = 1;$$

$$W_{Load} + W_{RSRQ} = 1 \text{ and}$$

$$0 \leq W_{RSRQ}, W_{Load} \leq 1 \quad \forall i$$

where T_i is the amount of traffic carried by CC_i and x_{Load_i} and x_{RSRQ_i} are the metrics associated to one UE and CC_i. Two weights (W_{Load} and W_{RSRQ}) are defined representing the relevance of the available bandwidth and the signal quality for the traffic split, respectively. Both proposed weights are configured by the network operator and sum up to one, so that increasing the relevance of the CC load (W_{Load}) implies reducing the relevance of the signal quality (W_{RSRQ}).

Figure 3-31 represents the maximum 5th, 50th and 95th percentiles for the UE download throughput in two cases: 1) following a homogeneous traffic distribution and 2) using the proposed traffic distribution when the network is configured with the weights W_{RSRQ} and W_{Load} which provide the maximum values of both 5th and 50th percentile. Therefore, this figure shows the benefit achieved by users when the proposed traffic distribution solution is used. In general, it can be seen how all throughput percentiles are higher when the proposed traffic distribution is used. Moreover, the achieved gain is greater as the percentile is lower. Thus, the proposed traffic distribution attains a better usage of network resources that it is more beneficial for users with worse conditions.

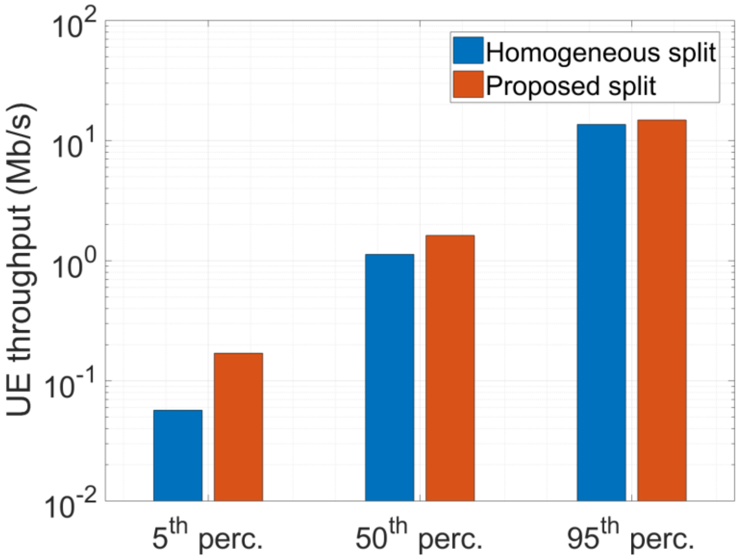


Figure 3-31 5th, 50th and 95th percentiles for the UE throughput following a homogeneous traffic distribution and using the proposed traffic distribution when the weights configuration is optimal

Additionally, a new technique called Uneven Traffic Split has been developed. This technique is based on a rule-based system and consists on an autonomous algorithm that carried out a traffic split among the serving CCs in a RSRQ and available PRBs basis. Therefore, each of the serving CCs may provide a different amount of resources to the UE. In this study, Uneven Traffic Split aims at maximizing the MOS of the four simulated services: eMBB FTP service (FTP), Non-Real Time Video (NRTV), Real Time Video (RTV) and Web Browsing (WB). Moreover, this technique could be combined with CC Manager so that a higher increase in MOS may be reached. In this way, Figure 3-32 represents the gain in the MOS achieved in three cases facing a baseline where the CC selection is carried out in a RSRP basis and the traffic split among the serving CCs is homogenous. In the first case, Uneven Traffic Split and the CC selection in a RSRP basis are carried out. The next case uses the CC Manager and a homogeneous traffic split among the serving nodes. Lastly, the combination of Uneven Traffic Split and CC Manager is carried out in the third case. While the CC Manager provides a higher gain when compared to Uneven Traffic Split, the figure shows that the combination of both techniques allows to achieve a further increase in the MOS. Moreover, it should be noticed that the gain is roughly higher for non-real time services.

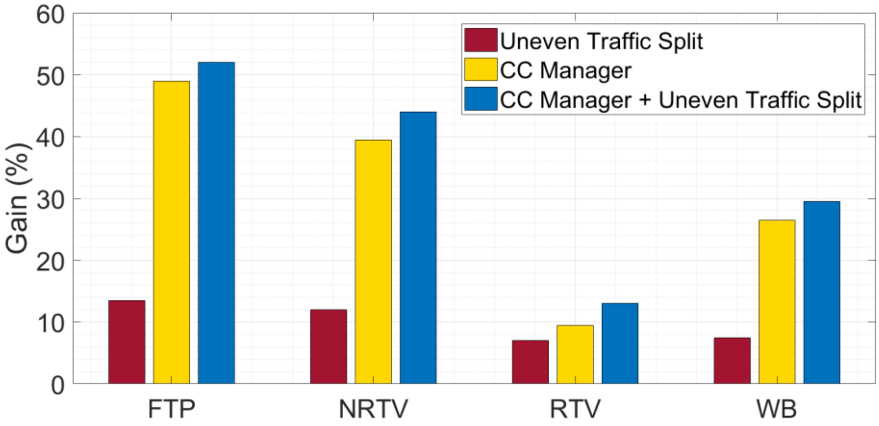


Figure 3-32 MOS gain when applying the Uneven Traffic Split solution

In summary, we draw the following observations and recommendations:

- An automatic allocation of CCs to a UE relying on a rule-based system whose rules use both quality (RSRQ) and load metrics has been proposed. The proposed solution provides up to 100% throughput gain for the users experiencing the worst throughput values over the state-of-the-art RSRP-based solution, and up to 75% gain at the peak throughput.
- The previous CC manager algorithm has been improved by including QoE metrics as input. This new algorithm obtains an average MOS gain of more than 70% compared to baseline situation.
- A heterogeneous traffic distribution among the nodes serving a multi-connectivity capable UE has been assessed. Results show that this new technique allows QoE metrics for different services to be increased when compared to homogeneous split traffic flows. In particular, if the traffic split is done according to the available bandwidth and the signal quality (RSRQ), the results show a significant gain of more than 50% in the MOS of FTP service and more than 10% for real time video services regarding the baseline.

3.3 Summary

Table 3-3 Summary of key recommendations and benefits in terms of multi-service and context aware radio resource management optimization

Feature	Recommendation	E2E / KQI benefits
Delay optimal user and channel scheduling	Myopic scheduling policies (e.g. max weight which is throughput optimal, etc.) are suboptimal if the KPI/KQI is the average delay.	30% improvement of average packet delay (service integrity) for eMBB users. Asymptotic optimality (in terms of average achieved delay) can be reached for some scenarios.
Dynamic resource allocation for URLLC services	For the cases where the CSI is not available at the transmitter, dynamic RB allocation for URLLC services can provide substantial gain as compared to static policies (i.e. that always allocate a fixed number of RBs to increase robustness).	50% improvement of service reliability and delay for URLLC services (if CSI is not available), higher supported throughput
Spatial and temporal availability of URLLC services	Jointly consider space & time in resource allocation for availability/ reliability improvement.	Improved service availability/ reliability considering a URLLC service area (e.g., a factory floor).
Configured grants for periodic non-synchronous uplink URLLC traffic	Use proposed scheme to continually adjust BS' estimate of traffic periodicity and time of arrival of next packet.	The possible reduction of overhead allows to support a higher number of URLLC uplink users. For 1 ms latency requirement, 3 or 30 times more UEs using 10 ms or 100 ms cycle time can be supported, respectively. In other words, the service accessibility KQI is significantly improved.
Downlink multiplexing of eMBB and URLLC service classes	Use preemptive scheduling for cases with moderate number of gNB antennas (supported in NR Rel-15). For deployments with eight or more gNB antennas per cell, use the developed MU-MIMO null-space preemptive scheduling solution.	Improved KQI service integrity for both the URLLC and eMBB users, higher supported offered load per cell, so clear E2E performance benefits. The MU-MIMO null-space preemptive scheduling solution offers 60% capacity improvement.
C-RAN multi-cell scheduling of URLLC traffic	Take advantage of low complexity fast dynamic multi-cell scheduling of URLLC users (from one cell per TTI). Based on UEs reporting at most CSI measurements from the strongest 2-3 cells within a received power window of 6 dB. Segmentation of URLLC is beneficial if limited to at most one UE per cell per TTI. Simpler DPS URLLC multi-user resource allocation is found to also be attractive.	Significant reduction (factor of two) of the experienced latency at low outage levels. Higher offered (30%-60%) aggregated URLLC traffic can be tolerated without violating the latency and reliability requirements. Improved KQI service integrity. The simpler form of centralized DPS URLLC scheduling offers up to approx. 30% latency reduction.

C-RAN multi-cell scheduling of eMBB traffic.	Centralized multi-cell scheduler for handling eMBB traffic is more suitable for high dense scenarios with very large number of users connected.	Improved average throughput per cell around 150% DL and 260% UL in a manhattan scenario; and around 77% DL and 260% UL in canonical scenario. Ensuring all users obtain resources.
Multi-legs configuration	It is proposed to estimate the achievable latency/reliability from the different "legs" which can be used to improve the reliability when sending up to 2 copies at a time.	Achieves optimal use of resources from the different component carriers, thus improving the KQI service integrity (i.e. tail of latency and reliability KPI) of URLLC traffic.
Operation for PDCP duplication for URLLC service classes.	It is proposed to use a) a novel duplication status UE report to timely acknowledge reception of a PDCP packet to multiple nodes, thus enabling in-network discard; b) selective duplication upon failure of the first packet transmission, thus avoiding duplicating when not necessary; and c) differentiating scheduling at the secondary node to avoid queuing delay for other traffic than URLLC.	Significant reduction of resources used for duplicated packets can be achieved, which results in substantial improvement in the KQI service integrity (i.e. tail of latency and reliability KPI) of URLLC traffic.
Automatic allocation of CCs to a UE.	The usage of a rule-based system whose rules use quality (RSRQ), load and quality of experience metrics is recommended.	Results show up to a 100% gain for the users experiencing the worst throughput values over the state-of-the-art RSRP-based solution, and up to 75% throughput gain at the peak throughput. Also, an average MOS gain of more than 70% compared to baseline is obtained.
Smart traffic distribution in a multi-connectivity scenario.	Exploit the benefits of an uneven traffic flow split in a multi-connectivity enabled scenario, by making the split rate depend on both the signal quality (RSRQ) reported by the UE over the CC(s) being used and their current load. A rule-based system based on these metrics is recommended.	Results show a significant gain in both the 5th and the 50th percentiles for the UE throughput when compared to a situation of equally split traffic flows. Using the Uneven Traffic Split method, it is possible to achieve gains between 50% and 10% of the MOS for different services regarding the baseline.

4 Spectrum, connectivity and mobility optimizations

In this chapter, the enhancement of the E2E performance is addressed by means of a number of mechanisms which rely on dynamic spectrum management, mobility optimizations and device-to-device (D2D) connectivity schemes. This is illustrated in Figure 4-1. In the first section, dedicated to spectrum-based optimizations, some general considerations of 5G NR bands operation are studied. Also, new bands expected to be used for 5G are evaluated from a service mapping view. Later, different requirements to be reached by the unlicensed bands operation and service provision are analyzed. Finally, a standard of independent unlicensed operation (Multefire) is optimized focusing on URLLC services. The second section focuses on novelties regarding mobility and load balancing in NR bands, based on QoE and context (i.e., information coming from events that, despite being external to the cellular network, have an impact on its performance, as the user location and speed). MEC-approaches in cellular V2X communications (C-V2X) are proposed as well as resource slicing schemes. The third section reflects new techniques that will allow the improvement in D2D communications based on the optimal management of resources for eMBB and mMTC with special emphasis on energy consumption. Relay techniques are also explored for mMTC and eMBB, including the dynamic management of resources.

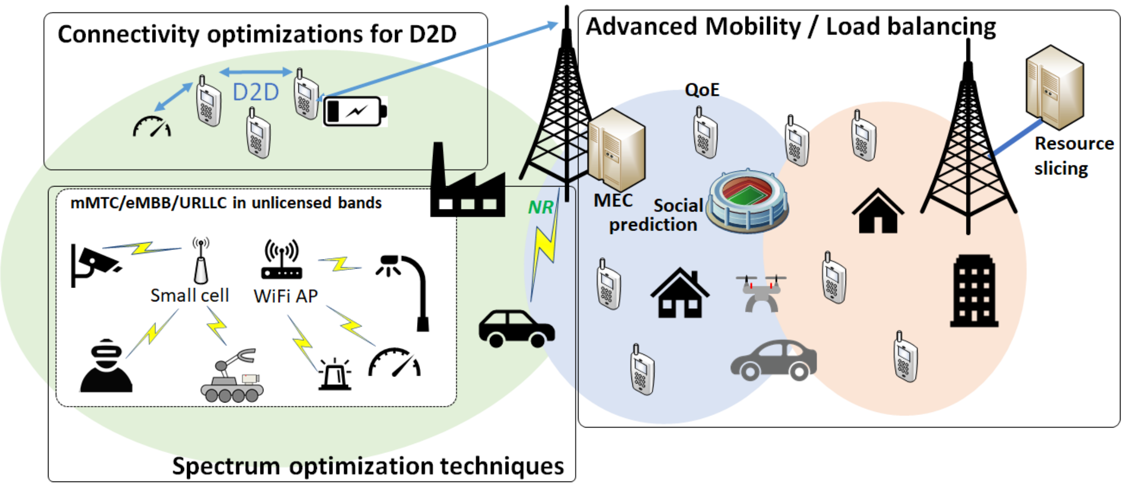

Connectivity optimizations for D2D
D2D
Advanced Mobility / Load balancing
QoE
Resource slicing
MEC
Social prediction
NR
mMTC/eMBB/URLLC in unlicensed bands
Small cell
WiFi AP
Spectrum optimization techniques

Figure 4-1 Spectrum, connectivity and mobility optimizations overview

4.1 Spectrum optimization techniques

4.1.1 mMTC operation in unlicensed bands

This subsection addresses some of the available frequency bands that may be used for mMTC unlicensed communication. The main focus is on the European regulations, but some options for other regulatory bodies like in USA and Japan are also treated, as some of these may potentially be considered for later expansion of the European regulatory domain. One of the main requirements for mMTC is reliable coverage; e.g. for sensors and actuators to always be connected. Achieving this in unlicensed bands will be more challenging compared to licensed bands due to restrictions on maximum transmit power output, limitations on spectral emission, channel access restrictions, and in general uncontrollable interference environment. Some of these aspects may potentially be addressed by system design mechanisms such as retransmission attempts and diversity techniques, but still, there is no guaranteed access to the spectrum resources. In the following review the pros and cons of using different unlicensed bands for mMTC are outlined.

5 GHz band: This band is denoted as band 61 in [EU2017], and is commonly called the “5 GHz WiFi band”, but is generally open for use by any radio technology that adheres to regulatory requirements. The band spans the frequency range from 5725 to 5875 MHz, allowing for a total of 160 MHz spectrum for unlicensed use. In this band, the maximum equivalent isotropic radiated power (EIRP) by any device shall be limited to 25 mW. To ensure efficient coexistence, the devices operating in this band must perform channel sensing prior to transmitting in the spectrum. Most devices operating in this band are operating according to the harmonized standard from ETSI [ETSI5GHz], since this will ensure general compliance with the regulatory requirements. By operating according to [ETSI5GHz], devices will have their transmission occupying at least 80% of a frequency channel, which will be operating in multiples of 20 MHz. Additionally, the transmit power is bounded by the spectral power density of, at most, 10 dBm per MHz. At least three radio standards are targeting operation in this band. These are based on specifications from IEEE, 3GPP and the MulteFire alliance. The IEEE specifications are described in the 802.11ac standard available in [IEEE11ac], the 3GPP specifications [3GPP] are based on licensed spectrum support to implement licensed assisted access (LAA), while the MulteFire specifications are described in [MFA]. The main challenge for operations using mMTC devices in band 61 is the bandwidth utilization requirement, which will normally cause devices to use high data rate or very short channel utilization time, which in turn will render low volume data communication inefficient.

2.4 GHz band: This band is denoted as band 57a or 57c [EU2017], and is also commonly denoted as the “2.4 GHz WiFi band”, but similarly to the 5 GHz band, any radio technology that follows the regulatory requirements is allowed to operate on it. The band spans the frequency range from 2400 to 2483.5 MHz, allowing for a total of 80 MHz spectrum for unlicensed use. The target device types for band 57a are short range devices, while the target device types for band 57c are radio local area network devices. Both device types have access to the same spectrum, so it is important that coexistence mechanisms are implemented accordingly. The coexistence mechanisms used for this spectrum range from frequency hopping as implemented for Bluetooth to channel sensing as implemented for IEEE and MulteFire specifications. When accessing the spectrum as a wideband radio device, the regulations allow for 10 dB higher transmit power compared to short range devices which are allowed to transmit at maximum 10 mW EIRP (Effective Isotropic Radiated Power). The main challenges when operating in this frequency band are the reduced transmit power and the relative busy spectrum due to a high amount of technologies operating on it. mMTC operation in this spectrum may be challenged by high interference levels from the common technologies operating in this band, which are IEEE 802.11b, IEEE 802.11g, IEEE 802.11n, Zigbee, Bluetooth.

865-870 MHz band: Recently, the European Commission released bands 47b and band 54 for short range devices. A total of five relative narrowband channels of up to 200 kHz are available with a maximum transmit power of 500 mW EIRP. Since transmissions in these bands allow for higher transmit power, devices operating here will have possibility for extended coverage, and coexistence is obtained through duty cycle limitations, which is limited to 2.5% for devices, while network access points (base stations) can have a duty cycle up to 10%. Operating with high transmit power in a low frequency band on a relative narrowband channel suits the requirements for mMTC communications but may be challenged by the other technologies that are allowed in this spectrum. Especially the RFID interrogator channels are coinciding with the band 47b, and may create areas with coverage problems.

US 902-928 MHz band: The FCC in US allows for unlicensed transmission in the 902-928 MHz band. The FCC regulations allows for using frequency hopping in the spectrum with at least 25 different hopping frequencies used with bandwidth of between 25 kHz and 500 kHz. For frequency hopping operation, the devices may transmit up to 36 dBm EIRP under specific configurations. The operation in this band is subject to regulations also requiring pseudo randomness for the frequency hopping. The operators using the band are not allowed to apply any coordination mechanisms to ensure interference free carriers.

<u>Japan 1.9 GHz band:</u> The regulatory bodies in Japan have released a frequency band normally used for DECT systems to be used by other radio technologies. The target technology for this band is a modified TD-LTE system, where the coexistence is implemented in the base station side, and devices to be used in this spectrum can be regular TD-LTE UEs that are frequency tuned to this specific band. As this regulatory region is quite small, it may not be interesting for mMTC operation, but the tendency from regulatory bodies in Japan is interesting since an existing dedicated band is released to be used by other technologies. We could envision similar things happening for the DECT band in Europe.

To summarize, the possible bands for unlicensed operation for mMTC usage are captured in Table 4-1. Limitations of duty cycle may not be limiting for low volume traffic that can respect the requirements.

Table 4-1 Summary of possible unlicensed bands for mMTC operation

Unlicensed band	Standardization/ regulation bodies	Technical features	Comments
5 GHz	ETSI, FCC, IEEE, 3GPP, MulteFire alliance	BW >= 20 MHz PSD and Tx power limited LBT based coexistence	New spectrum, wide spectrum available. Potential coverage problems that may be addressed by repetition in time. Hidden nodes may be a challenge in case of unbalanced link budgets.
2.4 GHz	ETSI, FCC, IEEE, MulteFire alliance	BW between 1-20 MHz Tx power limited to 100 mW LBT based coexistence or Frequency hopping for interference averaging	Challenged by the reduced transmit power and the relative busy spectrum due to relatively high number of technologies operating in this spectrum.
865-868 MHz	ETSI, IEEE	BW<200 kHz Tx power limited to maximum 500 mW Duty cycle based coexistence	Well suited for low volume communication Might suffer from uncontrolled interference from RFID interrogator channels.
902-928 MHz	FCC, IEEE	BW between 195-570 kHz Frequency hopping or digital modulation Tx power limited to 36 dBm	Only available in USA region, but still well suited for low-medium volume communication.
1900 MHz	Land Wireless Communication Committee, Japan	BW = 5 MHz Tx power limited Slow LBT based coexistence	Limited global availability. System operation is relying on existing systems having good coexistence mechanisms.

4.1.2 Radio resource allocation strategies for services mapping

The World Radio Conference (WRC) has started multiple discussions to identify new frequency bands that should be used in 5G. Among the main frequency bands that are specified in [3GPP38.101] we find frequency bands below 6 GHz called **FR1**, where typically the bands n77

(3300 – 4200 MHz) and n78 (3300 – 3800) have been selected for 5G. Bands above 6 GHz, called **FR2**, are also proposed; typically bands n257 (26500 – 29500 MHz) and n258 (24250 – 27500 MHz) seem to be the good candidates for 5G. FR1 and/or FR2 intra and inter-bands aggregation are possible for bandwidth extension. [3GPP38.133] specifies procedures and requirements (handover, cell selection, measurements …) when aggregating FR1 and FR2 frequency bands. In continuity of what was already described in [ONE18-D31], the main goal of this study is to benefit from the inter-band aggregation of FR1 and FR2 frequencies for increasing the network densification, especially when considering ultra-dense urban environments. The main question we try to answer is: ***"What is the best resources allocation strategy for the users being covered either by FR1 or FR2 or both?"*** The study takes into account the cell load and also the type of service to transmit (eMBB and URLLC). We propose a new strategy that optimizes the cells load repartition by proposing a "pre-scheduler" upstream on top of ofclassical state-of-the-art schedulers that ensures the KPI requirements to be achieved. In this study, the KPIs we try to improve are: the spectral efficiency, the latency and the QoE of the network.

4.1.2.1 Key performance indicator

The study carried out consists in improving the QoS offered by multi-cell networks. Different KPIs have been used to evaluate the overall QoS provided in the network:

- The spectral efficiency is the quantity of transmit information on a resource unit corresponding to a portion of bandwidth. Depending on the scheduler and its ability to use the channel knowledge, the mean spectral efficiency may differ. A basic scheduler will not consider the channel state and therefore not profit from its variation, in opposite to an opportunistic scheduler, which uses the CSI in its scheduling decision.
- The bandwidth usage ratio is an indicator of the mean ratio between the number of RBs used and the total number of available RBs. A high ratio reduces the flexibility of the system to sustain unexpected traffic peaks.
- As a QoE indicator we consider the PDOR (Packet Delay Outage Ratio) characterized by the ratio between the number of packets out of delay and the total number of arrived packets. It is an essential indicator in a multi-services scenario, especially for applications with high time constraints.

4.1.2.2 Multi-cell aggregation pre-scheduling algorithm

We consider a gNB-µGNBµgNB scenarioµGNB where UEs can be simultaneously served by the two types of cells. UEs inside the µgNB are also inside the gNB, but the opposite is not necessarily true. Consequently, UEs inside both cells must be scheduled in priority by the µgNB in order to offload gNB bandwidth which has already several UEs to schedule, which are exclusively covered by it. An optimization problem appears when the µgNB is overloaded. Indeed, in this case, the gNB cell has to manage the exceess traffic from µgNB. This traffic is generated by UEs more or less far/profitable from the gNB. In this context, we propose a pre-scheduler called Multi-cell Aggregation Scheduler (MAS). This pre-scheduler is able to adequately arrange UEs. Its objective is that the µgNB first manages the less profitable UEs in term of throughput for the gNB. The excess traffic coming from the µgNB is more likely generated by close/profitable UEs for the gNB decreasing its overall impact.

The MAS solution compares the UE's CSI (in terms of CQI) in each cell to determine what cell is the most suitable for each UE such as:

$$MAS_k = \overline{M_k^{micro}} - \overline{M_k^{macro}}.$$

MAS_k represents the value for UE k determined by the difference between its mean spectral efficiency on the µgNB ($\overline{M_k^{micro}}$) and its mean spectral efficiency on the gNB ($\overline{M_k^{macro}}$). The MAS_k corresponds to the benefit of schedule with priority the UE in the µgNB. The more the MAS_k is, the higher its negative impact on the gNB is.

The pre-scheduler sortssort UEs by their MAS_k values. In the case where µgNB is overloaded, we calculate the limit ratio of UEs that must be scheduled in priority by the µgNB. Accordingly to this limit ratio of UEs, µgNB manages the percentage of UEs that have the higher MAS_k letting the gNB schedule the more advantageous UEs:

$$Limit = \frac{\overline{M^{m\iota cro}} * a^{micro}}{\sum_{k=0}^{K} B_k}$$

where K is the number of UEs, $\overline{M^{m\iota cro}}$ the mean spectral efficiency of the micro cell measured from previous allocations, B_k the buffer occupancy of a UE k and a^{micro} the number of RBs available in the µgNB. In the below sub-section, this pre-scheduler in association with an opportunistic scheduler is evaluated and compared to classical scheduling techniques.

4.1.2.3 Simulation assumptions

Typical 5G scenario assumptions have been implemented for performance evaluation at system level. We assume in our set-up that a centralized controller is implemented to coordinate both gNB and µgNB. Our deployment scenario is based on one of the twelve scenarios that have been introduced in [3GPP38.913]. The dense urban microcellular deployment scenario focuses on macro gNB with micro, high user densities and traffic loads in city centres and dense urban areas. The key characteristics of this deployment scenario are high traffic loads in outdoor coverage. In our network deployment we have considered 3 micro per macro,with the macro as control unit .

Table 4-2 Simulation parameters

Parameters	gNB	µgNB
Cell radius (m)	500	200
Carrier frequency (GHz)	3.7	26
Total bandwidth (MHz)	100	400
Transmission Power (dBm)	44	33
Resource Block	Numerology 0	Numerology 0

In order to test our "new band allocation” scheduler, we propose a scenario where all UEs can be served either by the µgNB first or the gNB or simultaneously if more resources are needed. No interference from µgNB or gNB is experienced by users. Their traffic model is variable and the time constraint depends on the service used. We consider two services, with time constraints of 10ms and 100ms. Our MAS solution is associated with two schedulers MaxSNR [WX06] and WFO[2] [GB09]. WFO is based on the PDOR that represents the ratio of packets arrived out of delay. This scheduler is designed to quickly react to this indicator in order to fairly maintain the level of QoE for all users. The objective is to determine if service-oriented schedulers conserve their characteristics when associated with our MAS solution. We compare the combination of the MAS solution with MaxSNR or WFO with the same scheduling methods without MAS.

[2] Weighted Fair Opportunistic

4.1.2.4 Performance Results

Figure 4-2 and Figure 4-3 show that when the number of users increases, the opportunistic schedulers exploit the multi-user diversity in order to increase the spectral efficiency. The MaxSNR provides the best results in terms of spectral efficiency (Figure 4-2 and Figure 4-3) stretching the system's congestion limit (Figure 4-4). However the system's capacity increase is made at the expense of its fairness (Figure 4-6 and Figure 4-7), resulting in a poor QoE quantified by the PDOR KPIs (Figure 4-5).

WFO improves the MaxSNR fairness (Figure 4-6 and Figure 4-7) making a trade-off with the spectral efficiency (Figure 4-2 and Figure 4-3).

Adequately associating the users to the most profitable cells, the MAS pre-scheduler combines successfully the advantage of both state-of-the-art schedulers. Adding MAS pre-scheduler to WFO slowly decreases its multiuser usage in the µgNodeB but is widely compensated by a more advantageous users allocation to the macro. MAS+WFO has 20% more users capacity than WFO alone in this context while guarantying its fairness abilities.

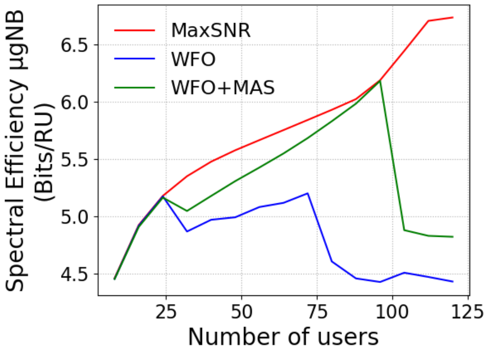


Figure 4-2 Spectral efficiency in the gNB

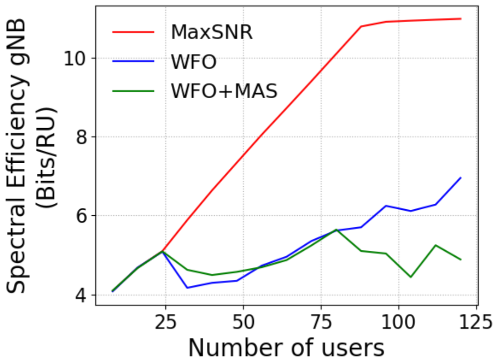


Figure 4-3 Spectral efficiency in the gNB

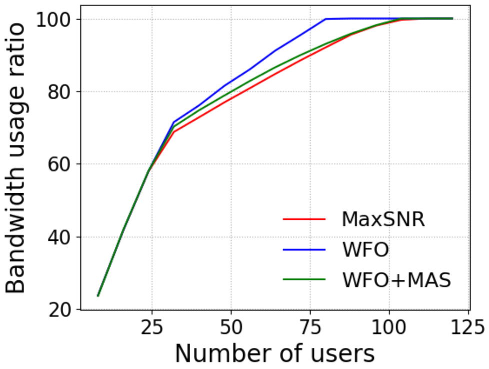


Figure 4-4 Percentage of bandwidth used

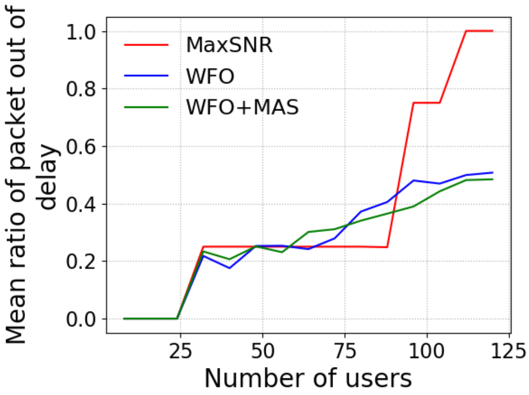


Figure 4-5 Mean PDOR in the system

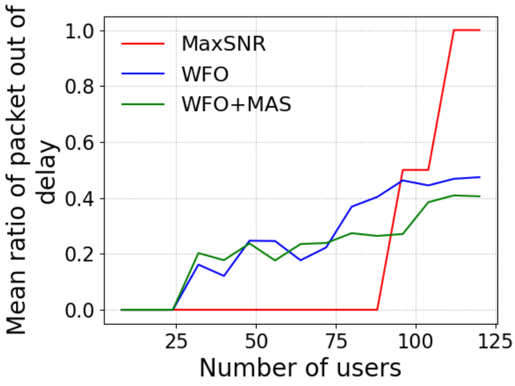


Figure 4-6 Mean PDOR for users close from gNB

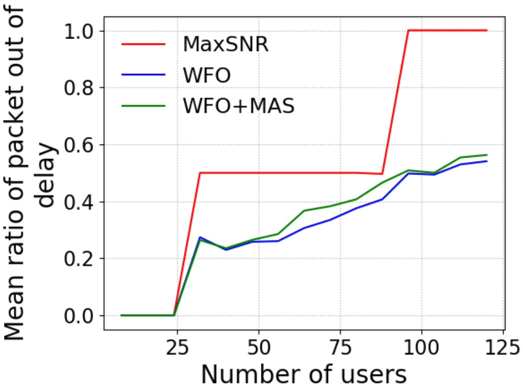


Figure 4-7 Mean PDOR for users far from gNB

4.1.2.5 Conclusion

The proposed multi-cell pre-scheduler is a design that greatly improves the performance of the WFO scheduler in terms of spectral efficiency and system users capacity, while preserving its fairness abilities. MAS+WFO is able to ensure multi-services differentiation without impacting the throughput, reaching same performance than the well-known MaxSNR resource allocation scheduling. Our approach is recommended on multi-services multi-cells context, mixing different service types with different QoS constraints such as eMBB and URLLC services.

4.1.3 Dynamic spectrum aggregation for 5G new radio

The concepts of dynamic spectrum aggregation in NR have been detailed in [ONE18-D31]. We summarize them in this section with the intention to achieve a complete view of the relevant research for this section.

In the 3GPP NR system, dynamic bandwidth aggregation/adaptation is supported to adapt UE transceiver bandwidth according to momentary traffic demand so as to optimize the UE power consumption. Specifically, according to [3GPP38.211], for a given UE, a configured DL (or UL) BWP may overlap in frequency domain with another configured DL (or UL) BWP in a serving cell.

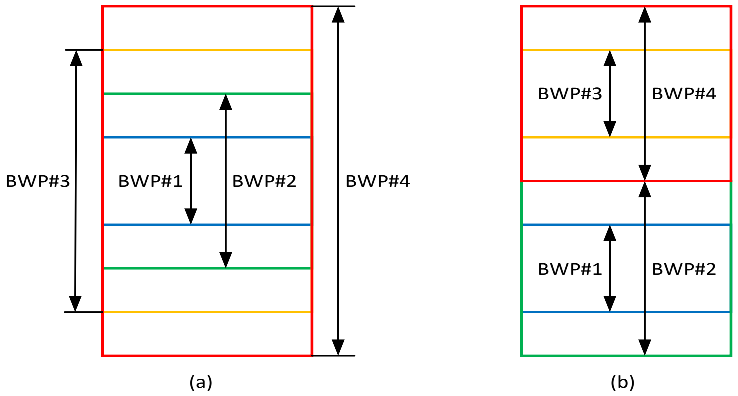


Figure 4-8 Multiple BWP configuration for NR UE, (a) bandwidth adaptation, (b) bandwidth adaptation and load balancing

For each serving cell, the maximal number of DL/UL BWP configurations is 4 DL and 4 UL BWPs, and 4 DL/UL BWP pairs for paired and unpaired spectrum, respectively. For paired spectrum, DL and UL BWPs are configured independently in Rel-15 for each UE-specific serving cell. For unpaired spectrum, a DL BWP and an UL BWP are jointly configured as a pair, with the restriction that the DL and UL BWPs of such a DL/UL BWP pair share the same center frequency but may be of different bandwidths in Rel-15 for each UE-specific serving cell.

With the support of multiple DL/UL BWPs with potential different central frequencies and bandwidths, dynamic bandwidth adaptation and load balancing can be readily realized for NR UEs. By virtue of the method described in Figure 4-8 (a), UE can be configured with four BWPs with the same central frequency but different bandwidths; dynamic switching between 4 configured BWPs can adapt the UE receivedto receive bandwidth according to the instantaneous traffic throughput demand. With another BWP configuration method illustrated in Figure 4-8 (b), two pairs of BWPs are located in different central frequencies, and two BWPs in each pair share the same central frequency and have different bandwidths. Dynamic BWP switching in this case would enable bandwidth adaptation for the UE and load balancing between different spectrum parts in the system. Moreover, for the UE, dynamic BWP selection can be performed according to the UE short-term channel condition so that bandwidth efficient transmission can be achieved.

4.1.4 KQI based analysis of eMBB services in unlicensed bands

The use of unlicensed bands is one of the most promising features envisaged to increase capacity in 5G. However, this poses multiple challenges associated to the operation with coexisting networks, such as WiFi. Previous coexistence analyses have been focused on the user-plane data-related transmissions and mainly based on abstract models [BCG+17]. Meanwhile, the effects of the in-band signaling defined by the standards [3GPP36.205] have been mainly disregarded, particularly for ultra-dense scenarios. An example is the case of Discovery Reference Signals (DRS) which, according to our previous results [ONE18-D31], have a significant impact on service performance. In Annex G we present an optimization to improve service performance.

In this way, we have analyzed the performance of FTP Model I with a Poisson arrival distribution according to certain traffic intensity given by lambda λ (transfers per second) as an eMBB service representative. Besides, its KQI integrity indicators i.e “**File Transfer Average Throughput**” measured in Mbps (FTAT) service and “File transfer Delay or FTD (s)” have been quantified.

In order to complement the optimization shown, an assessment has been made of the different possibilities of accessing the non-licensed band channel in order to establish a framework for the optimization of different services based on its KQI for different conditions of the indoor scenario.

According to 3GPP standards, LBT mechanisms have been considered for Downlink Shared Channel unlicensed access. LBT relies on clear channel assessment procedure (CCA) with energy detection (ED) threshold to sense the channel state for a defer period whether any signal is present above a certain threshold (regardless of its kind). In case the channel is detected free then the LAA station is allowed to transmit for a Maximum Channel Occupancy Time (MCOT) otherwise enhanced CCA (eCCA) is activated and has to wait for a backoff period of time determined by a Contention Window (CW) measured as a number of time slots.

Table 4.3 3GPP LBT priority classes

Priority class (p)	CWmin	CWmax	MCOT
1	3	7	2 ms
2	7	15	3 ms
3	15	63	8 ms
4	15	1023	8 ms

As shown in Table 4. within this category, four priority classes have been defined according to different values of MCOT and CW size. According to the standard, lower values have higher priority. This channel access configuration remains in the standard for future specifications applicable to non-licensed bands in New Radio (NR-U) [3GPP38.889] including signaling. These priorities, similar to the mechanism Enhanced Distributed Channel Access (EDCA) established from 802.11e in WiFi, serve to provide QoS, which in this case will be measured based on the KQIs. A real time (Video Streaming) and a non-real time (FTP) service has been evaluated for different load conditions. For real time Video Streaming the following Integrity KQIs have been considered respect to the ones proposed for standard (adaptive) Video Streaming.

Table 4.4 Proposed real time Video Streaming KQIs

	Real Time Video streaming
KQI- Integrity	Frame Loss Ratio (%) [NGMN16]
	Cumulative Jitter (s) [ETSISTQ] Frame e2e delay (s) [ETSISTQ]
	Video Peak Signal to Noise Ratio PSNR (dB) [NGMN16] Rate-based Objective Mean Opinion Score ROMOS [AKR+14]

Peak Signal Noise Ratio (**PSNR**): is a measure that compares the video signal pixel by pixel and frame by frame. It is totally ignorant of the spatial and temporal relationships of the pixels and, as a result, very different distortions that clearly result in very different quality perceptions obtain similar PSNR values. Nevertheless, it is a first valid approach to measure QoE.

Rate-based Objective Mean Opinion Score (**ROMOS**): is an objective QoE video performance indicator [AKR+14] which includes streaming network parameters such as the frame loss rate.

The average KQI for each of the commented services has been measured in different conditions, both users per cell and for traffic intensity per user (measured in terms of FTP arrival rate) as well as including or not WiFi coexistence. Based on the results obtained as well as the study of the signaling, we have selected the optimal priority and signaling. All this is summarized in the table below.

Table 4.5 Unlicensed bands KQI Assessment summary

Service	**Network Load (FTP)**	**WiFi Coexistence**	**Best LBT CAT4 Priority class**	**Recommended Signaling**	**Avg KQI achieved**
FTP Model I **(non real time)**	Low ($\lambda<0.5$)	Yes	Priority 1	C-DRS	FTT: 128.76 FTD: 17.31
		No	Priority 4	DRS DMTC 40 ms	FTT: 125 FTD: 5,9
	High	Yes	Priority 4	C-DRS	FTT: 119.27 FTD: 19.85

	(λ >2.5)	No	Priority 1	DRS DMTC 40 ms	FTT: 123.21 FTD: 18.31
Video Streaming (real time)	Low (λ <0.5)	Yes	Priority 1	C-DRS	ROMOS: 4,86
		No	Priority 1	DRS DMTC 40 ms	ROMOS: 4,95
	High (λ >2.5)	Yes	Priority 4	C-DRS	ROMOS: 3,85
		No	Priority 4	C-DRS	ROMOS: 4,12

Based on the assessment in Table 4.5, dynamically aggregating a carrier with an optimal combination of channel access configuration depending on the type of service demanded by the user, we can get a substantial KQI performance improvement (up to 40% best case Annex G 7.7) with respect to standard (non service-oriented) configuration with fixed priorities and signalling use. Such results have been combined with an enhanced fairness values towards WiFi coexisting devices.

4.1.5 Unlicensed standalone operation

Standalone operation of 3GPP-based radio access technologies in unlicensed spectrum have recently attracted the interest of large-scale enterprises looking for wireless solutions that can provide reliable access to private networks globally, and without the need for expensive licensed spectrum. Support for LTE standalone operation in the 5 GHz unlicensed band has been introduced with the first release of MulteFire [MFA1.0], [RKF+18]. More recently, 3GPP has started a work item on New Radio-based access to unlicensed spectrum (NR-U), with standalone operation in the 5 GHz (and 6 GHz) unlicensed spectrum being one of the target deployment scenarios [3GPP_NRU]. However, so far both MulteFire (MF) and 3GPP standardization bodies have explicitly targeted enhanced mobile broadband use cases with only minor focus at introducing support for high-reliable and low-latency communication using these unlicensed bands.

An important aspect when considering operation in the 5-6 GHz unlicensed spectrum is that the most restrictive regulatory requirements mandate clear channel assessment based on LBT. LBT is a contention-based protocol that allows devices to use the same radio channel without pre-coordination. Transmission by a device on the radio channel is conditional on the device sensing the radio channel below an energy detection threshold. If the channel is occupied, the device is not allowed to transmit. Also, a device is only allowed to occupy the channel for a limited duration of time before it shall perform a new LBT procedure. Clear channel access based on LBT obviously represents an additional challenge when targeting highly-reliable and low-latency communication in 5 GHz unlicensed spectrum.

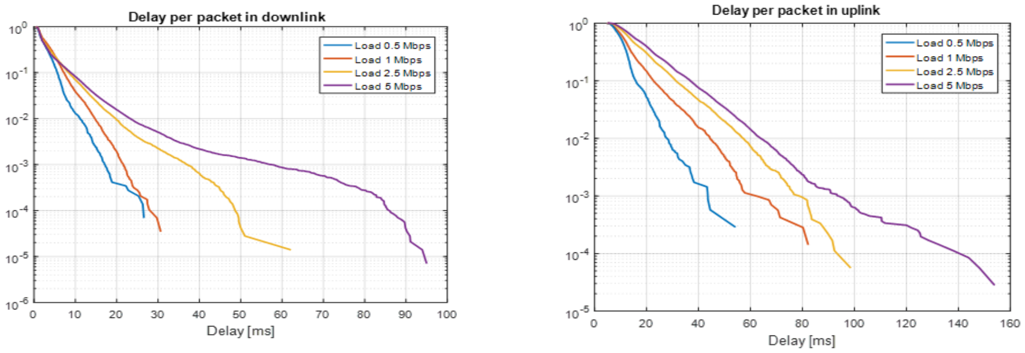


Figure 4-9 Distribution of the packet delay in downlink (left) and uplink (right) and under different load conditions.

We have started our analysis by assessing the reliability and latency performance in the 5 GHz unlicensed spectrum using a system level simulator compliant with the MulteFire radio specifications, and in line with the 3GPP indoor scenario for LAA coexistence evaluations in [3GPP36.889]. The analysis has been conducted for a single operator deployment, as would likely be the case for e.g. factory cases or larger enterprise installations. A dynamic traffic model corresponding to FTP model 3 in [3GPP36.889] is applied. Data packets of 50 Bytes are generated by each user according to a Poisson arrival process. The total network load is adjusted by varying the arrival rate of the Poisson process. Simulations assumes a 80:20 downlink (DL):uplink (UL) traffic ratio. As expected, the initial results in Figure 4-10 indicate that LBT has a significant impact on the experienced latencies on the radio interface, especially when estimating the maximum latency that can be guaranteed with a certain reliability. However, even if the most stringent URLLC requirements defined for 5G use cases can clearly not be achieved with LTE system operating in the 5 GHz unlicensed spectrum, these preliminary results indicated that use cases requiring one-way radio latency in the order of 50 to 60 ms with 99.9% reliability can be supported under specific load conditions.

LBT has a significant performance impact, especially in the uplink direction of transmission, due to the hidden node-problem. This is because, in a scheduled system, the access point contends for the medium also on behalf of the mobile terminals, while the interference conditions at the access point and at the mobile terminals can be relatively different. Uplink LBT failure is also impacting the downlink performance, since the stations can be hindered from transmitting uplink control information such as HARQ feedback. This can cause the access point to either unnecessarily retransmit packets, or to reduce the reliability of transmission by assuming the packets were successfully received.

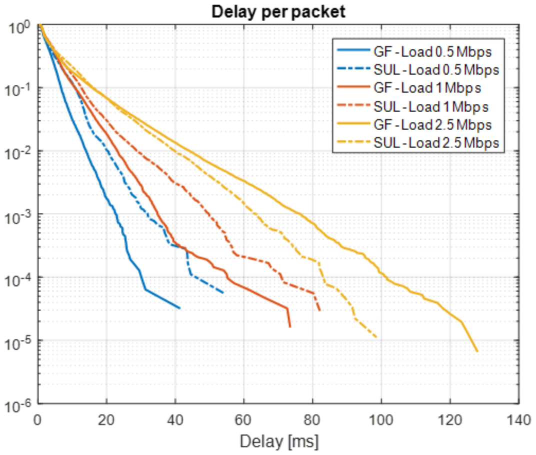


Figure 4-10 Distribution of the packet delay in uplink with scheduled uplink (SUL) and grant-free (GF) uplink.

First, the performance of grant-less uplink (GUL) transmission in unlicensed spectrum is evaluated [Note: GUL is aka Grant-Freee (GF)]. Scheduled uplink suffers from the inherent delays associated with the need to send a scheduling request and receive a scheduling grant prior to an uplink transmission. Therefore, grant-less uplink transmission schemes, where terminals can autonomously initiate transmission in uplink, provide significant advantages in terms of latency, especially in low load conditions. This effect is even more noticeable in unlicensed spectrum, due to the additional delays associated with the LBT procedure, which is required prior to any transmission on the channel. However, when the load increases, and for high reliability requirements, GUL starts to suffer from excessive collisions due to uncoordinated access to the uplink resources, and performance degrades behind that of scheduled uplink (Figure 4-10). Additional performance results and related details of the conducted studies are available in two conference papers; see [MRF+18a] and [MRF+18b].

To reduce the impact of uplink LBT failures on the downlink performance, we also investigated solutions that allow the network to provide the terminals with up to K additional opportunities for transmitting the HARQ feedback during each access point-acquired channel occupancy time (COT). If LBT fails at the uplink control channel occasion (PUCCH), the UE can attempt transmission of HARQ feedback at the first additional opportunity, and so on. Another way to alleviate the impact of uplink LBT failures when providing HARQ feedback is to ensure that the gap between the end of a downlink transmission and the start of the uplink transmission carrying the HARQ feedback is below 16 μs. In this case, the regulatory requirements allow the mobile terminal to initiate transmission without performing LBT. The impact of these enhancements on the downlink latency and reliability performance is illustrated in Figure 4-11. To observe a noticeable performance improvement, in low load conditions it is enough to increase the number of HARQ transmission opportunities per COT, while as the load increases, it becomes necessary to disable LBT for the transmission of HARQ feedback.

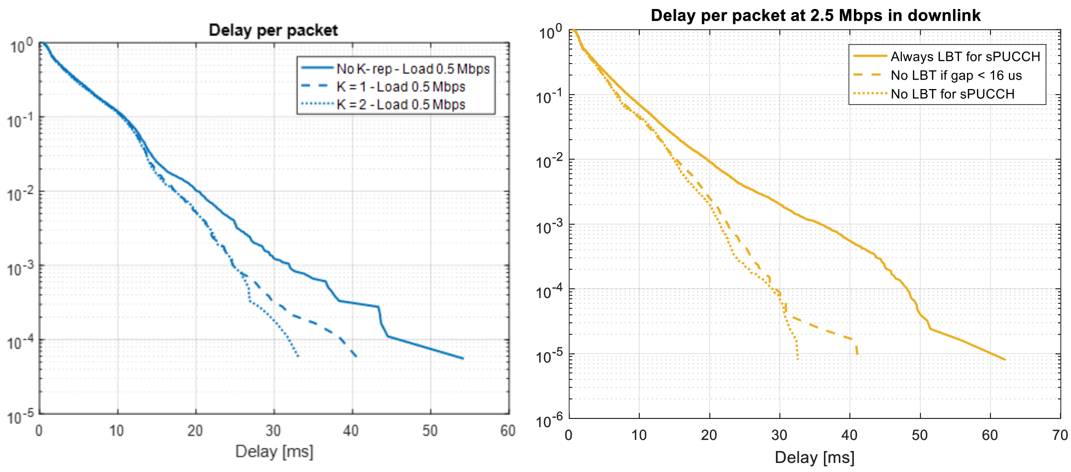


Figure 4-11 Distribution of the packet delay in downlink with multiple HARQ transmission opportunities per COT (left) and by disabling uplink LBT for the transmission of HARQ feedback (right).

In summary, we draw the following observations and recommendations for the studies based on the MulteFire system model (see also additional details in [MRF+19]):

- Achieving low latency communication in the 5 GHz unlicensed band is challenging due to the LBT procedures. As the offered load increases, our extensive performance results show that up to 40% of the total one-way packet delay budget is consumed by LBT.
- Using grant-free for uplink transmissions: 25% latency improvement at low to medium loads. No benefit at high offered loads.
- Enabling K-repetition for ACK/NACK feedback: Approximately 20% latency improvement at low to medium loads. No benefit at high offered loads.

- Omitting Cat 1 LBT during DL-2-UL transition: Up to 55% latency improvement when the offered load is high.
- Our results show that, under specific load conditions and using the proposed enhancements, systems complying with the MulteFire standard specifications can support use cases requiring one-way radio latency in the order of 30 to 40 ms with 99.9% reliability.

Adopting the NR-U system model further improves the latency-reliability performance by:

- Supporting higher subcarrier spacing and, therefore, shorter symbol duration.
- Supporting shorter Time Interval Transmissions (TTI).
- Improving the processing capabilities at both the gNB and the UE side.

Moreover, NR-U brings additional features that aim to cope with the additional challenges presented in unlicensed operation:

- Multiple switching points: multiple switching points within the TDD frame structure after a successful LBT are allowed. Providing higher flexibility to the frame structure favors the continuous transmissions after a single channel access (Category 4 LBT).
- Higher flexibility at starting frame occasions: NR-U includes the concept of mini-slots. In this way, a node can in principle start transmitting right after finishing the LBT or at least in multiple occasions as compared to LAA/MulteFire. By using this approach, the risk of losing the channel while waiting until the starting symbol is reduced.

Therefore, NR-U offers significant latency/reliability benefits over LAA/MF. Specifically, at 99.99% reliability, a reduction of 58%, 73%, 73% and 62% in latency is achieved for system loads of 0.5 Mbps, 1 Mbps, 2.5 Mbps and 5 Mbps, respectively. Equivalent to latencies of 8-17 ms at 99.99% reliability.

4.2 Advanced mobility optimization and fast agile load balancing mechanisms

4.2.1 Basic 5G NR mobility trends and solutions

In this section we briefly describe some of the basic mobility enhancements for 5G NR, before presenting novel proposals in the subsequent sections for mobility related optimizations. While the basic RRC connected mode mobility functionality for LTE is based on network controlled handovers with UE assistance, additional enhancements are introduced for NR, addressing some of the challenges identified in [BMS+12a], [BMS+12b], [3GPP36.839], [PBM+13]. Especially, the asynchronous nature of the LTE handover functionality with random access (RA) at every cell change is known to cause undesirable interruption times of 20-30 ms and up to 100 ms (or even more) for some networks [GMP16a], [EE13]. Such interruption times are especially critical for URLLC type of applications. To overcome this problem for 5G, it has been proposed to adopt a time-synchronized and RA-less handover functionality [BPR+15] (as also studied in the FANTASTIC-5G project). The basic principle of RRC Connected mode synchronous RA-less handover between a source and a target cell is as follows: once the source cell receives a measurement report to trigger the handover, the source and target cells exchange information, and agree on the time of the actual handover. The handover command (RRC reconfiguration message) informs the UE of the exact time of the handover. The UE may continue to receive data from the source cell until the time of the handover, where-after it starts to receive immediately data from the target cell. The handover command from the source cell may also include an uplink pre-scheduling command for the UE to immediately transmit in the uplink towards the target at the time of the handover. The handover interruption time is thereby reduced to a fraction of a subframe (or TTI), accounting for the received time difference of the signals from the two cells.

In a more advanced version of the synchronous RA-less handover functionality (as studied also in FANTASTIC-5G, D4.2), the UE continues to listen to the source cell for a short time period after the handover time, while also receiving data from the target cell in parallel. The use of synchronous RA-less handovers, also helps to reduce the overall handover execution process. The faster handover execution translates to increased mobility robustness, as the system can react faster [BPR+15]. For scenarios with high-speed users such as the highway scenario studied in [GMP16a], the use of synchronous RA-less handover also offers significant reduction of the required RA resources. Finally, optimized usage of synchronous handovers for cloud RAN architectures were recently studied in [KGP17].

The concept of conditional handover has also been discussed in 3GPP as a potential improvement for NR. The main idea of conditional handover is to communicate with the serving cell early while the link is still strong and access target cell late when the link is sufficient. As studied in details in [MVL+18], conditional handover provides clear gains although care in parameter setting must still be taken to avoid excessive signaling increase (and hence still calls for efficient parameter optimization algorithms).

For cases with multi-node connectivity (MC), several mobility related enhancements have been studied (and introduced) for 5G NR, as compared to LTE-Advanced Dual Connectivity. Among others, the use of partly UE autonomous cell management to reduce the otherwise heavy RRC signaling overhead [GMP16a], [PBM+13] has been studied by giving the UE the responsibility of secondary cell / gNB management decisions. As an example, for the highway scenario studied in [GMP16a], it is found that the average number of required RRC signaling messages per UE per second is reduced from 4.9 to 0.35 by using UE autonomous secondary cell management. Moreover, the associated inter base station signaling is reduced by approximately 50%. Furthermore, use of enhanced MC for 5G NR also brings mobility benefits. As one such example [VML+18], using signaling radio bearer duplication, radio link failures (RLFs) could be eliminated under certain assumptions. Using such techniques in combination with NR MC with PDCP data duplication, helps fulfil the URLLC requirements (even for users' subject to mobility). However, as also mentioned in [VML+18], further studies of self-organizing frameworks to manage the new parameters for adding and removing gNBs have to be developed (see the related studies in Chapter 3.2) as well as more advanced mobility robustness and load balancing optimizations (as are the topics of the sub-sequent sections).

4.2.2 Context-aware proactive QoE traffic steering through multi-link management

This section aims at describing a set of mechanisms to proactively configure network configuration parameters related to traffic steering. To that end, predictive techniques are first used to forecast the network performance, so a proper network configuration can be afterwards applied to avoid an eventual performance degradation.

4.2.2.1 Social events information gathering, association and application to cellular networks

As described in D.3.1 [ONE18-D31, section 4.3.2], the activity related to social events (e.g. concerts, sport matches, parades, etc.) has previously focused on the mechanisms for automatic data gathering from Internet sources and its association to cellular network data (such as the base stations positions and orientation) [FSB+17] [FPS+18]. Mechanisms for cellular indicators (e.g. KPIs) forecasting have been also developed based on neural network based nonlinear autoregressive exogenous models (NARX), with a novel combination of both cellular and social events related data.

Evolving from that, the capability to relate the impact of previous events in past performance metrics (KPIs/KQIs) values of the network has been developed, as a required input to trigger optimizations able to avoid future forecasted degradations (see Figure 4-12).

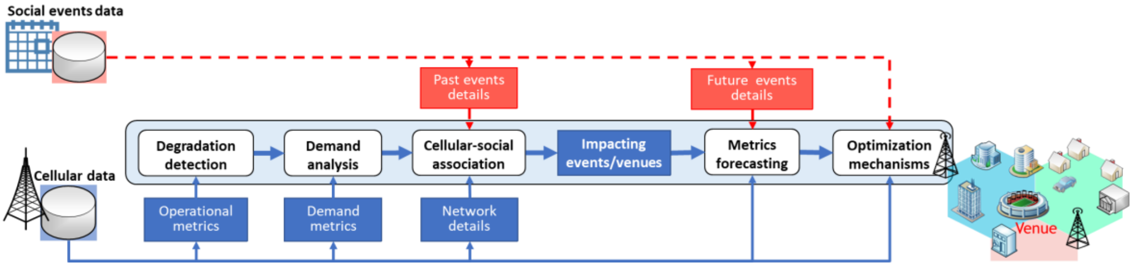
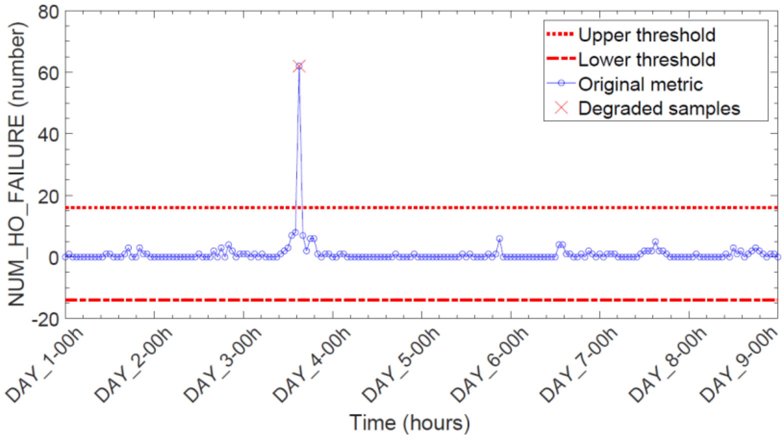


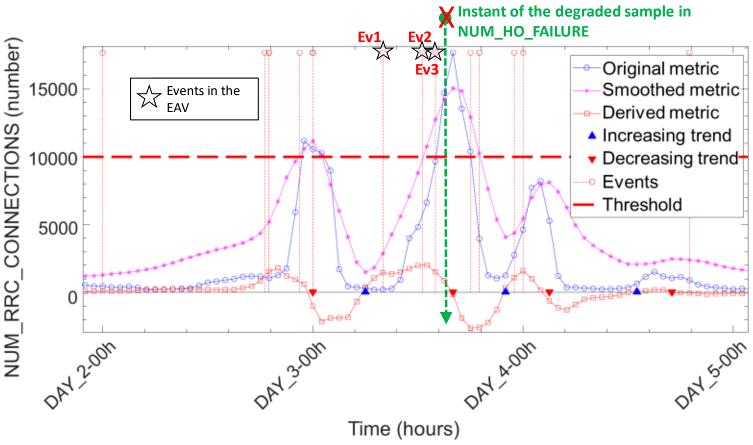
Figure 4-12 Scheme for social data application

Here, the initial approach [ONE18-D31, section 4.3.2] was based on fixed parameters for the definition of the temporal scope (named as “event associated window”, EAW) where a metric might have been impacted. However, this solution required manual definition of the number of hours before and after the start time of an event in order to define the search of possible impacted metrics.

Conversely, a new automatic method beyond SoTA has been implemented based on the analysis of the temporal evolution of the cellular metrics themselves. Here, cellular degradations are first detected automatically by the identification of outliers in “operational metrics”, this means, those associated with the well-being of the service provision (e.g. dropped connections, handover failures), as presented in Figure 4-13a. Here, both upper and lower thresholds are automatically generated, although one of them would be typically ignored depending on the nature of the metric (e.g. “number of failures” considered threshold should be the upper one, where throughput per user thresholds should be the lower threshold). If degradations are detected (by values crossing those thresholds), “demand metrics” (e.g. number of connections) variations are automatically analysed in order to find coincident profiles indicating the possible presence of an increased service demand generated by a social event. As shown in Figure 4-13b the identification of the increasing demand profile related to a social event is based on trend analysis. Here, the metric under analysis is smoothed to avoid spurious fluctuations. The zero crossings of its derivative serve to limit the EAWs where the possible past impacting events are investigated.



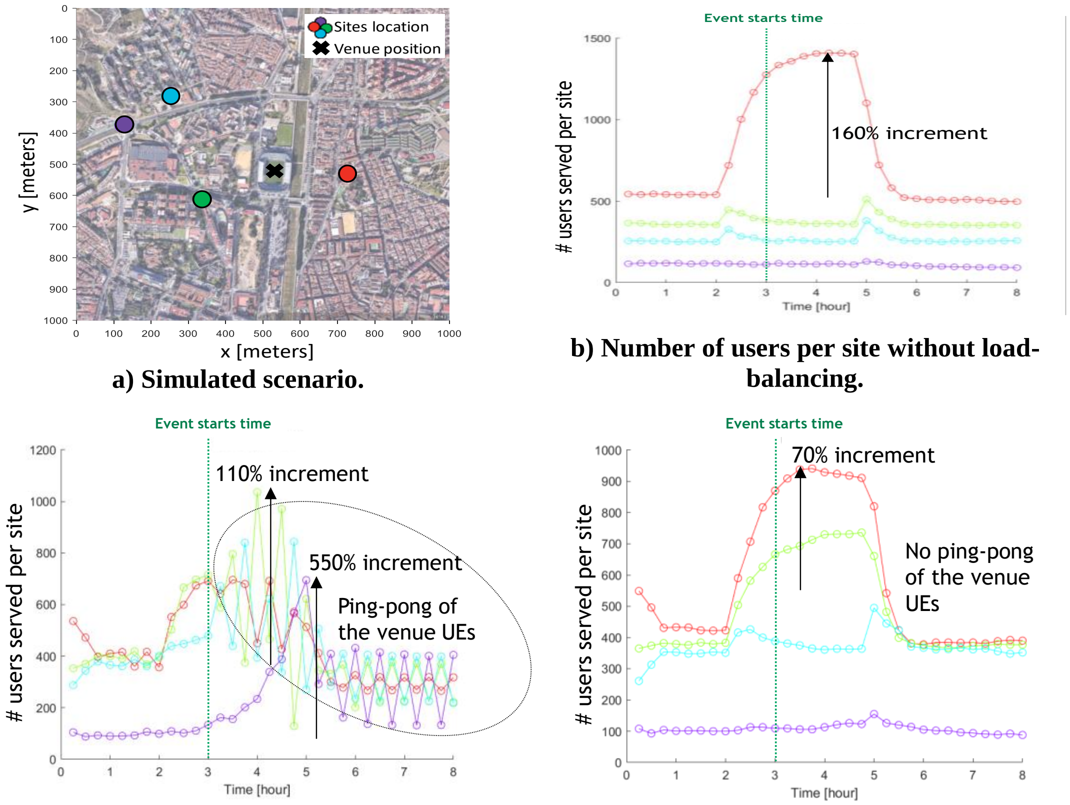
a) **Operational metric analysis.**



b) **Demand metric analysis.**

Figure 4-13 Example of the degradation detection and demand analysis for social-events

From this point, it is possible to identify the most likely events (and venues) causing the degradation through the "cellular-social association" (see Figure 4-12), where temporal, location and metric-based analysis can be used to identify the most impacting events/venues. This information can then be used for metrics forecasting (predicting the specific impact on the performance) and optimization of the network to avoid future degradations. For example, by identifying the venues whose events have a higher impact on the cellular network, it would be possible to minimize this impact in its future events. In comparison with pre-existent approaches these techniques allow for the automatic identification of social events as the cause of network degradations, which was not possible beforehand.



a) Simulated scenario.

b) Number of users per site without load-balancing.

c) Number of users per site with classic PTS.

d) Number of users per site with VA-PTS.

Figure 4-14 Social-aware optimization of the cellular network.

Additionally, social information has been also applied as an input for the optimization of the cellular networks. The available social information for the venues include the time of the expected event and specially, the precise location of its venue. The possibilities of this approach has been evaluated via simulation of a social event venue and the surrounding cells based on real sites deployment scenarios, as shown in Figure 4-14a. Different type of UEs (static, mobile and attending the event) have been defined and modelled based on real data of population density and venue capacity for a sport stadium. The event UEs are generated outside the scenario and move towards it before its beginning, leaving it after its end.

In this scenario, a power traffic sharing (PTS) load balancing mechanism [FAB+15] has been applied as baseline. This mechanism, implemented via a fuzzy logic controller, defines rules for the reduction of the transmitted power of overcrowded cells, leading to a reduction in their coverage areas and therefore the reduction of the number of users served by them, balancing the load of the network.

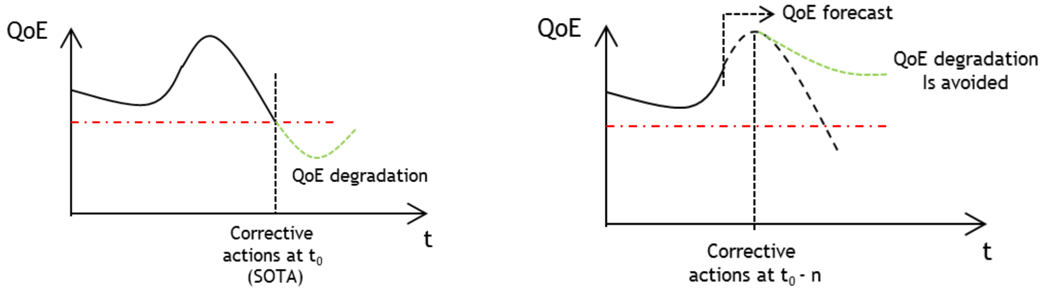
When a social event occurs, without load-balancing, there are huge variations in the spatial distribution of the users between cells (see Figure 4-14b). The described classic PTS load-balancing technique works typically well under uniform or semi-uniform slow-varying user densities. However, under non-uniform distributions, with dense and time-variant user spots, the controller behaviour will highly depend on the defined output values for the rules of the controller, leading to long to achieve stable values (if the modifications of transmitted power are too low) or overshoot, making the cell serving the venue to ping-pong between the closest pair of cells (see Figure 4-14c).

In order to overcome this issue, by knowing the position of the venue, a venue-aware PTS (VA-PTS) mechanism has been defined. This considers also the position of the next event venue, tuning the output of the previous PTS algorithm: where the venue is close to the centre of its main serving cell, power reductions can be high, as the venue would be still under its coverage and the balancing would be performed only based on dispersed cell edge users. However, if the users are close to the edge of the serving cell, the reduction of power is limited.

In this way, the proposed VA-PTS approach avoids the venue to ping-pong between different cells (see Figure 4-14d), highly reducing the total number of handovers and achieving also a 44% reduction (from 160% to a 70%) in the peak increment of users served by the site closest to the venue.

4.2.2.2 Network performance prediction enhancement through feature engineering

The aim of this research line is to predict the network performance, so that occasional performance degradations, eventually impacting the UE's QoE, can be forecasted and avoided by means of a proper proactive network configuration. Currently, optimization tasks only take place whenever a performance degradation occurs (Figure 4-15a), i.e., in a reactive manner. As a consequence, UEs may perceive such performance (QoE) degradation. By following the approach herein proposed, however, corrective actions could be taken some time before the actual degradation takes place (dotted green line in Figure 4-15b), attending to the forecasted behaviour (dashed black line) of the performance metric under evaluation rather than to its current value (solid black line).



(a) Current approach to deal with network performance degradations.

(b) Proposed framework to predict a network performance degradation and take a corrective action in advance.

Figure 4-15: Current and proposed framework to deal with network performance degradations.

To that end, a number of (output) network performance indicators are forecasted using both a predictive model and a number of (input) performance indicators. Besides, in order to reduce the operational complexity of the predictive model, as well as the necessity for a wide range of input performance data while preserving or even improving the prediction accuracy, feature engineering is used over the input performance indicators. That is, a set of machine learning techniques which, relying on a feature selection approach (such as a sequential feature technique [ROS11]), aim at identifying which performance indicators and which *lags* (the time difference between a predicted sample and an input sample) to use at the input to achieve a dimensionality reduction with the least loss of useful information.

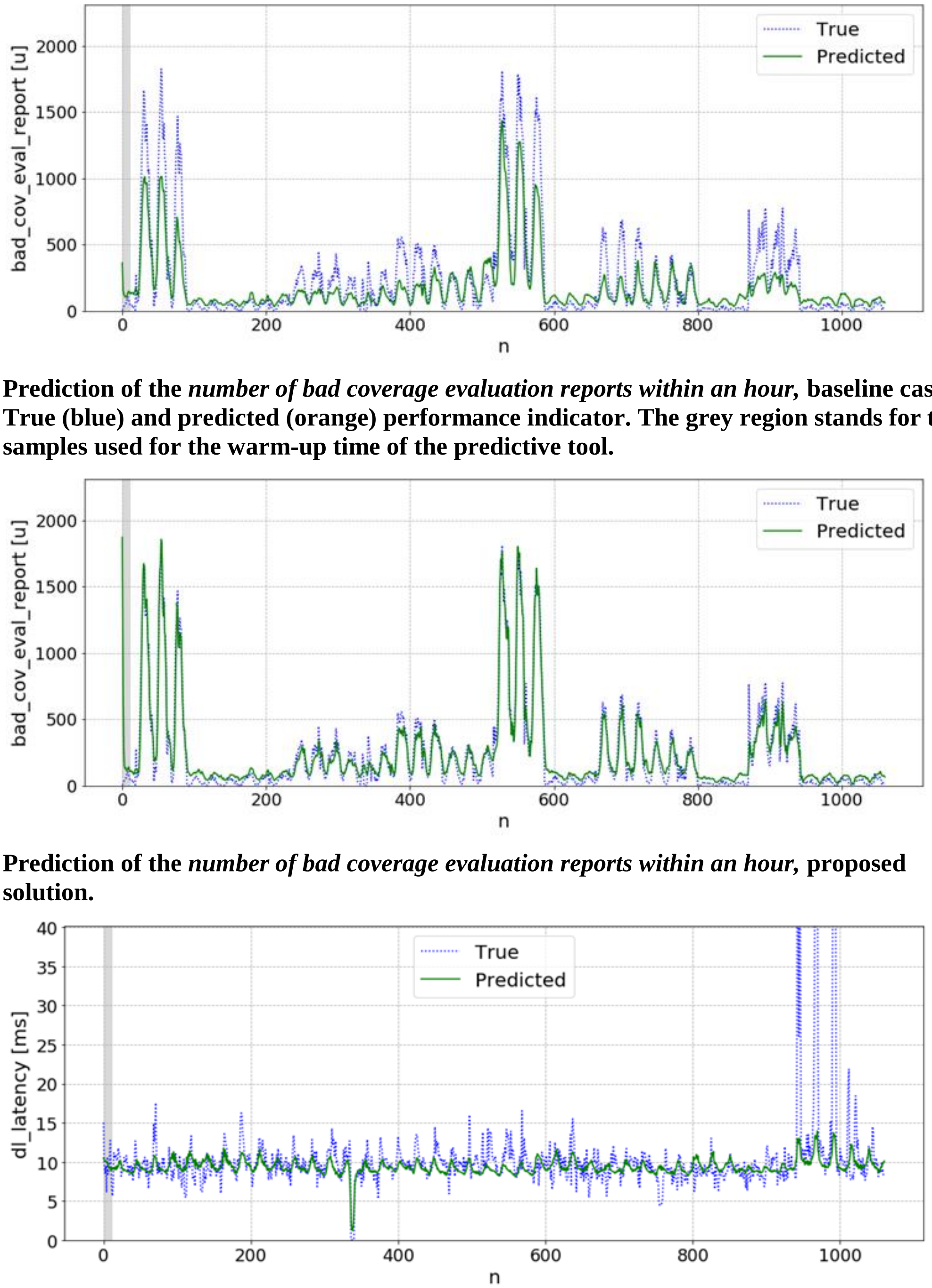
A proof of concept has been carried out to assess the feasibility of the proposed approach. The aim of this proof of concept is to assess the benefits in terms of computation time and prediction accuracy of using a feature selection tool prior to the algorithm for prediction. To that end, a database, made up of 43 cell-aggregated KPIs and approximately 30 thousand samples, gathered on an hourly basis from a live LTE network has been used. Each lag/KPI pair has been considered as a different performance metric, so that a feature selection technique can be applied to identify the pairs contributing the most to the prediction of the output KPIs. In this case, up to 24 lags in the past have been considered, leading to a total of 1032 lag/KPI pairs (see (1)). In this test, the KPIs to be predicted are the *number of times that a bad coverage evaluation report is notified within an hour*, the *hourly-averaged user downlink latency (in ms)* and *user downlink throughput (in Mb/s)*. The time shift to be predicted is 1 hour into the future ((t+1), in this dataset). The tool that has been used for prediction is a state-of-the art recurrent neural network, composed by several gated recurrent units (GRUs) [CMG+14]. Feature selection has been carried out by using a random forest regressor [BRE01]. This technique has been fed with the 1032 input KPIs; and a score, corresponding to their relevance for the output KPIs to be predicted, has been computed. The 20 highest scoring lag/KPI pairs have been consequently used as the input for the GRU network. The baseline for this proof of concept consists in using all the 1032 pairs as the input for the GRU network. The figures of merit used to compare the proposed approach and the baseline case are the computation time for the training and test phases of the prediction algorithm and the root mean squared error (RMSE) achieved in the test phase.

$$\begin{pmatrix} x_1(t) & x_1(t-1) & \cdots & x_1(t-24) \\ x_2(t) & x_2(t-1) & \cdots & x_2(t-24) \\ \vdots & \vdots & \ddots & \vdots \\ x_{43}(t) & x_{43}(t-1) & \cdots & x_{43}(t-24) \end{pmatrix} \tag{1}$$

In any of the subfigures of Figure 4-16, the prediction of a single KPI is evaluated over time. These subfigures show the predicted KPIs (in orange) and the true KPIs (in blue) along 1000 samples in the test stage for both the baseline case (Figure 4-16a,c,e) and following the proposed approach (Figure 4-16b,d,f). The figures of merit comparing both cases are summarized in Table

4-6. As it can be seen, making a proper selection of the inputs to be used by the forecasting method contributes to both noticeably reducing computation times and the resulting RMSE, which, in this case, has been lowered by a factor of 20. Since this feature selection will only be applied once, prior to the construction of the prediction model, its computation cost is negligible, compared to the resulting benefits of its use.

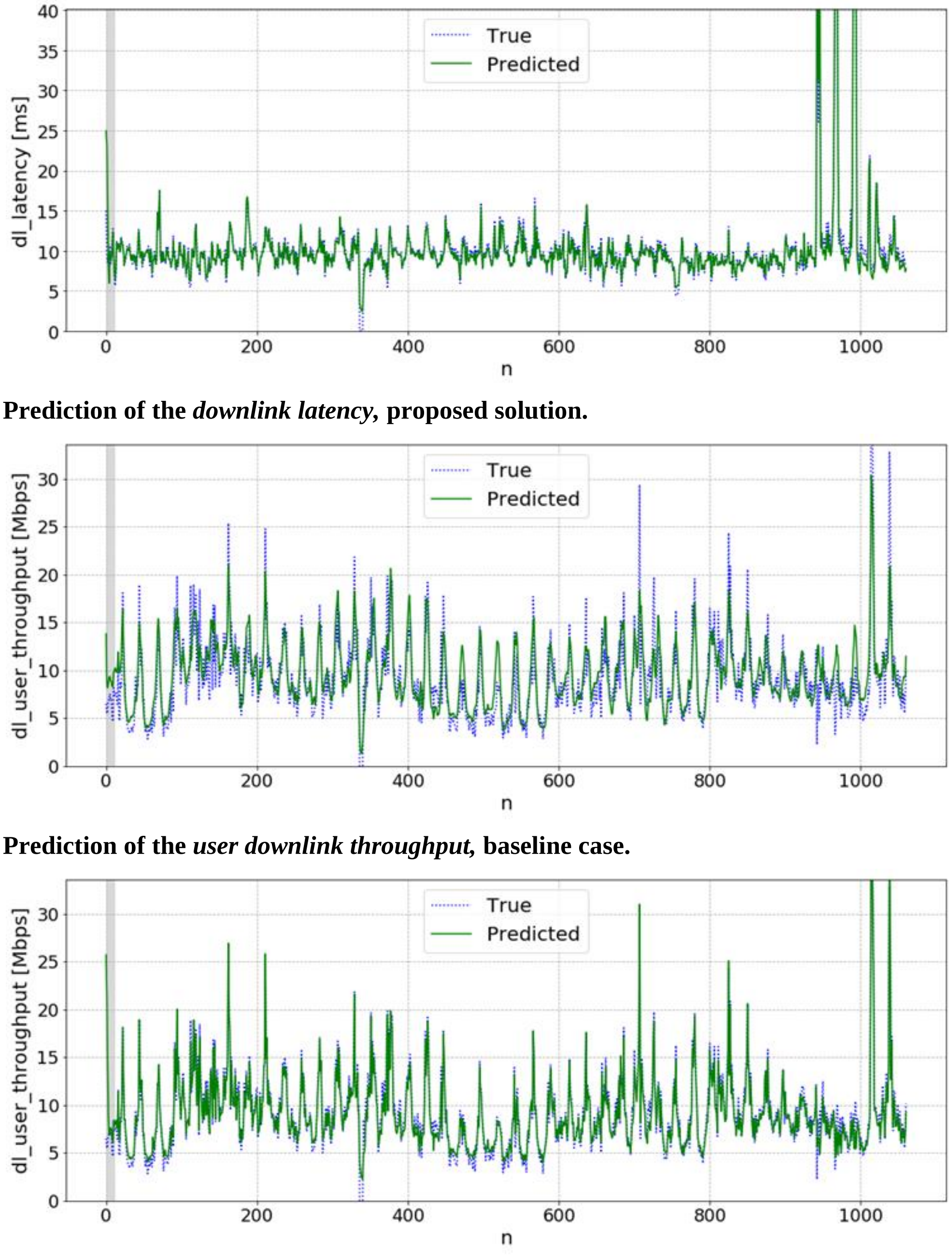
Thus, the proposed framework is shown to noticeably improve the prediction accuracy, while reducing computation times, paving the way for its integration in a tool for a proactive QoE management by only relating the predicted performance indicators to an eventual QoE metric by using expressions like [eqs. for KPI-to-QoE mapping from section 3.2.2].



(a) Prediction of the *number of bad coverage evaluation reports within an hour,* baseline case. True (blue) and predicted (orange) performance indicator. The grey region stands for the samples used for the warm-up time of the predictive tool.

(b) Prediction of the *number of bad coverage evaluation reports within an hour,* proposed solution.

(c) Prediction of the *downlink latency,* baseline case.



(d) Prediction of the *downlink latency,* proposed solution.

(e) Prediction of the *user downlink throughput,* baseline case.

(f) Prediction of the *user downlink throughput,* proposed solution.

Figure 4-16 Prediction results (test stage) on *t+1* for different KPIs over time (*x* axis).

Table 4-6 Figures of merit comparing the baseline case versus the proposed solution

		Baseline: 1032 lag/KPI pairs	**Feature selection + prediction: 20 lag/KPI pairs**
Training stage	**Processing time**	12’45’’	5’ 24’’
Test stage	**Processing time**	21’’	7’’
	RMSE	$3\cdot10^{-3}$	$1{,}43\cdot10^{-4}$

4.2.3 Network slice management based on mobility and traffic patterns

Due to mobility and service demand variations the traffic levels change in time and space and this would mean that network slices do not have to be always enabled and at every location all the time (which would lead to resource inefficiency). Therefore, network slice management can be extended to proactively create network slices according to required service demands and ongoing network characteristics. However, sometimes the requirements for the services that will be present in a specific area are not known in front of time and reactive approaches if adopted, can lead to degradations or inefficiencies (e.g. over-provisioning or high control latency in order to cover the required load).

In the first year of the project, our work focused on negotiation of resources and was presented in D3.1 [ONE18-D31], [WIN19]. In the proposed solution of the second year of the project, the user mobility is taken into account e.g. assuming that users are moving from area A to B (e.g. residential to business areas) during specific time periods (e.g. midday) which will lead to service demand variation calling for appropriate allocation of resources to business area ‘B’ for handling extra traffic from inbound users due to mobility. At the same time, allocation of resources to residential area A can be reduced during midday since less users will be there (due to their movement in the business area ‘B’). This scenario is presented graphically in Figure 4-17.

Mobility patterns are already used in latency, coverage and resource sharing related proactive solutions. Here we propose to use mobility patterns in network slice management solutions during the preparation stage of network slice lifecycle. Proactive management of slices will enable better control of which slices the network will allocate to UEs based on acquired data and targeting KPIs. It will also lead to elimination of potentially negative impact of transient situations and will lead to resource efficiency, optimized topology and specific configuration to serve various service requirements. Finally, it will allow for further differentiation of what slices may provide to slice customers or end users, with features like enabling networks e.g. for factory automation, ultra-reliable and ultra-low latency and more.

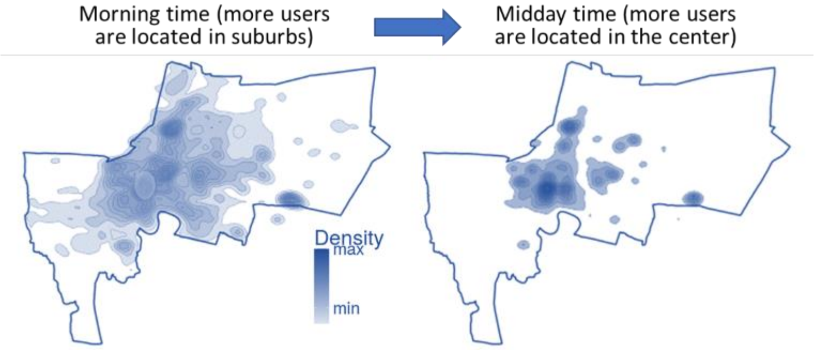
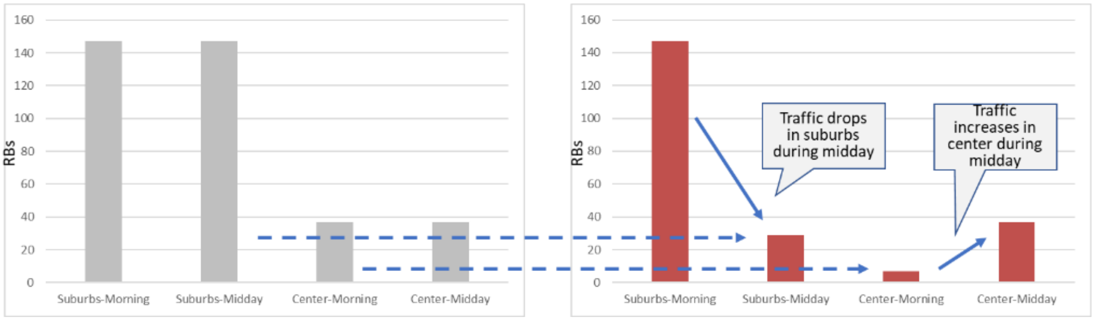

Morning time (more users are located in suburbs)
Midday time (more users are located in the center)
Density
max
min

Figure 4-17: User densities changing in different times in suburbs and center [CDH17]

The proposed mechanism which allocates resources in different areas and times of day due to user mobility are located in Communication Service Management Function (CSMF) of the 3GPP Network Slice related management functions ([3GPP28.801], Sec. 4.10). CSMF is responsible to translate the available mobility per service type patterns of the users into slice related requirements. Then CSMF communicate with the Network Slice Management Function (NSMF) ([3GPP28.801], Sec. 4.10) for the actual realization of Network Slice Instances (NSI) on the 5G network (creation, management and orchestration of slices). The proposed solution is executed during the Preparation phase of the NSI lifecycle ([3GPP28.801], Sec. 4.1). The proposed solution contributes to the preparation of network slice templates and of the necessary network environment which are used to support the lifecycle of NSIs, as well as the pre-provision of the network slices.

The following results illustrate the potential savings in resources due to traffic drop in suburbs during midday and also the reduced traffic in center area during morning (compared to midday). The provided results are based on analytical study and it is evident that resource efficiency of over 50% is anticipated due to the mobility of users from suburbs to center (and vice versa). In the first case (Figure 4-18a) the maximum required resources are allocated both to the slices located in centers and suburbs. In the second case (Figure 4-18b) the proposed solution proactively allocates resources to the slices by estimating the traffic based on mobility patterns. Main improvements can be summarized to the following:

- Savings in resource utilization are foreseen in suburbs in midday compared to the overprovisioning case. Specifically, there is a drop in usage from around 140RBs (assuming a traffic of 3,000Gbit/km^2/day) to less than 40RBs (due to less demand during midday in suburbs because many users have moved to the center as the user density suggests in the previous figure).
- Also, resource efficiency (compared to over-provisioning case) is anticipated in the morning around the city center, since less demand is experienced. As such, RBs in the morning for the center area drop from around 40 RBs to less than 20 RBs (due to the fact that less users are located around the center).
- Overall, through these results we evaluated the impact of mobility of users at different times of day and different place. Due to mobility and service demand variations the traffic levels change in time and space and this would mean that network slices do not have to be always enabled and at every location all the time and this will lead to resource efficiency.



(a) (b)

Figure 4-18: Allocated resources in different areas and times of day due to user mobility using: (a) resource over-provisioning and; (b) proposed solution

4.2.4 Latency-centric solutions for V2X communications: UE connectivity state handling and network architecture perspectives

V2X communication paves the way for a drastically improved road safety and driving experience via reliable and low latency wireless services [5GAA18]. However, V2X applications are part of the URLLC service type, and, as such, they have very stringent requirements in terms of E2E latency. Hence, it is important that the current V2X system is thoroughly examined to unearth possible latency bottlenecks, before proposing enhancements and alternatives, in order to meet the strict delay requirements. In this section we tackle the issue from two different angles: we first assess possible benefits of exploiting the different RRC states for V2X applications to increase the number of devices served in the cell, subject to delay requirements posed by these applications. Then, focusing on cellular V2X (C-V2X) communications and concentrating on the vulnerable road user (VRU) V2X use case, we present an alternative approach to the current, remote cloud-based network architecture that facilitates multi-access edge computing (MEC) infrastructure to decrease the E2E latency of the VRU signal destined to a cluster of vehicles.

4.2.4.1 RRC State Selection for URLLC V2X

To meet the E2E latency requirements of V2X applications, it would be desirable that vehicles have an always-on connection with the cellular infrastructure. However, due to the expected increase of the number of both V2X and non-V2X devices that are served in each cell, an always-on connection might not always be possible, but the load demands, along with the delay requirements could be met by exploiting the different RRC states.

According to [VAR93], road capacity (Annex 7.1.1) may be increased by the use of tightly spaced platoons. Therefore, we assume that vehicles organize themselves in platoons and select a platoon leader. We also assume that vehicles in platoons are synchronized (i.e., all vehicles in the same platoon transmit at the same time), but different platoons are not synchronized. We assume two different modes of communications:

1. Communication via the Uu interface (direct network communication)
2. Communication via the PC5 interface (communication via sidelinks)

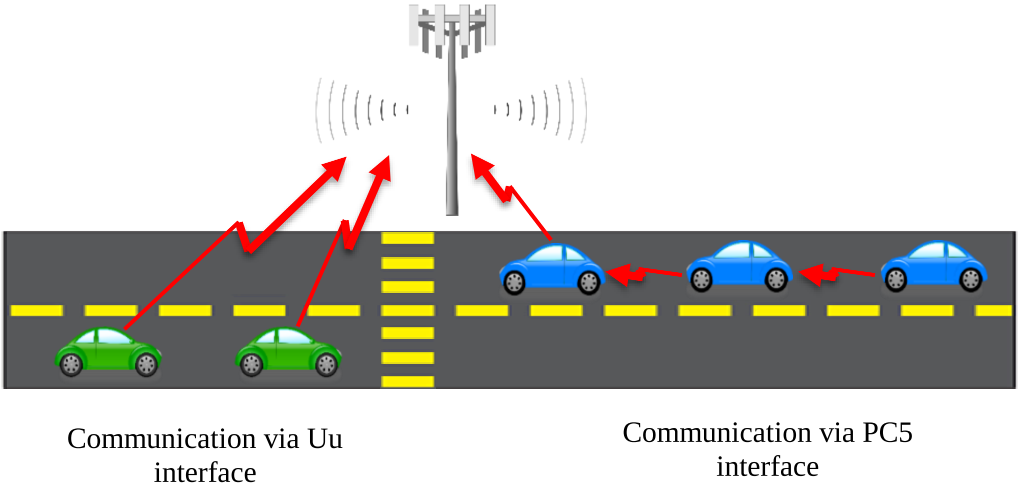

Communication via Uu interface
Communication via PC5 interface

Figure 4-19 Illustration of the two communication modes. The green cars form a platoon that communicates with the gNB via the Uu interface (direct communication). The blue cars form a second platoon where only the leader communicates with the gNB and intra-platoon communication is done via PC5 (sidelink) resources.

Regardless of the communication mode used, transmission of a message from vehicles which are in either the RRC Idle or RRC Inactive state, requires that the vehicles first transition to the RRC Connected state. This transition requires signalling exchange between the base station and the vehicles, and it must be considered as part of the total E2E delay.

We define the state transition latency T_{conn} as the average time required for a V2X device to switch from either the RRC Idle or the RRC Inactive state to the RRC Connected state, and the transmission latency T_{transm} as the time elapsed from the moment the V2X message is transmitted, until it reaches the final destination, including any network delays, assuming an already established connection. Therefore, the total E2E latency T_{total} for a device (either V2X or non-V2X) is defined as $T_{total} = T_{conn} + T_{transm}$. Based on the number of non-V2X devices served in the same cell, we assess the benefits of switching V2X devices to either the RRC Idle or the RRC Inactive state, while meeting the total E2E latency requirements of V2X devices.

For the first communication mode (i.e., via the Uu interface), we assume a 10 MHz channel, serving 1400 vehicles, split in platoons of 25 vehicles each, and a varying number of non-V2X devices. The T_{conn} for RRC Idle/RRC Inactive vehicles is given by Annex 7.1.2 based on the collision probability for each transmitting vehicle given by Annex 7.1.4. To better appreciate the effects of the state selection, we assume that $T_{transm} = 0$. However, as this would not be the case in a real-life scenario, the maximum observed transmission latency, T_{transm}, must be added. Therefore, based on the minimum state transition latencies in Figure 4-20(cases a and b), shows the state transition latency, $T_{\_trans}$ of V2X devices if they are switched to the RRC Idle or RRC Inactive states. We can see that when up to ~25000 non-V2X devices are present in the cell, then the RRC Idle state can be used by V2X devices, without violating the E2E delay requirement of 100 msec. When more than 25000 and up to ~38000 non-V2X devices are served in the cell, then the RRC Inactive state must be used, otherwise the E2E delay requirement of 100 msec may not be met. Serving even more non-V2X devices requires that vehicles must remain in the RRC Connected state to meet the delay requirements.

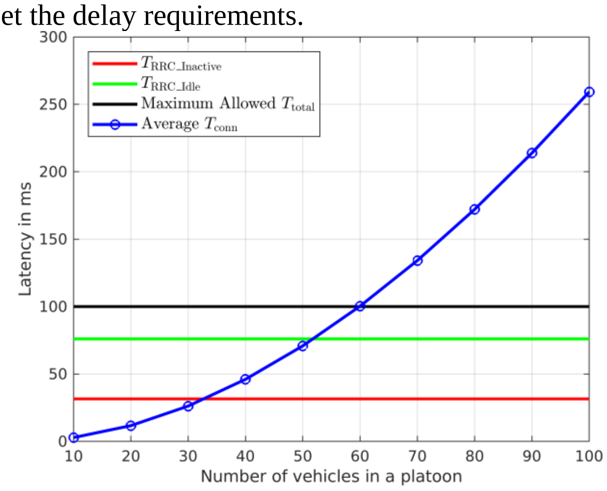


Figure 4-20 The figure depicts the estimated state transition latency per V2X UE for 1400 vehicles for a varying number of non-V2X devices in the cell.

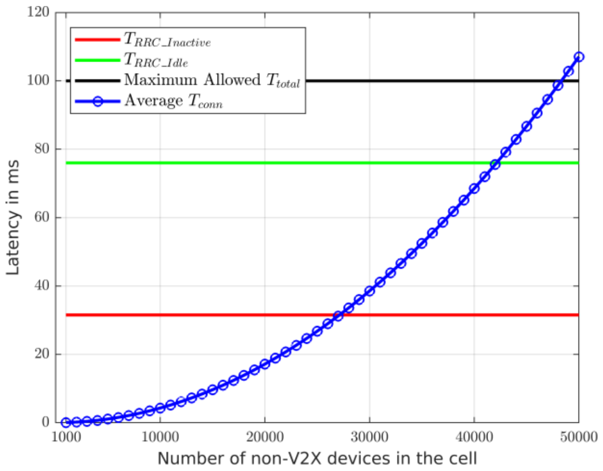


Figure 4-21 Correlation between the platoon size and the state transition latency for V2X devices.

For the second communication mode (i.e., via the PC5 interface), we also assume a 10 MHz channel, serving a variable number of platoons, each of which contains 25 vehicles, and 45000 non-V2X devices. We assume that vehicles in a platoon do not contact the platoon leader directly, but instead, each vehicle transmits its message to its immediate front vehicle, which then combines the received message with its own, before forwarding it to its immediate front vehicle, until all messages reach the platoon leader. As only the platoon leader is communicating with the base station in this mode, and the intra-platoon communication is done via sidelinks, we consider that the T_{conn} consists of the latency for message transmission within the platoons, and the state transition latency of the platoon leader which depends on the collision probability at the connection establishment phase. Furthermore, in this case T_{transm} refers only to the latency of the transmission of the final message by the platoon leader. Similarly, to the previous communication mode, we assume that $T_{transm} = 0$, in order to better appreciate the effects of using the two RRC states with different platoon sizes. Further, we assume that 1 resource block (RB) is allocated to the sidelink resources, and can thus not be used for normal communication procedures (i.e. connection establishment, data exchange). Figure 4-21 shows the correlation between the size of the platoon and the RRC state that the platoon leader can be switched to. It is apparent that, as the size of the platoon increases, the latency also increases.

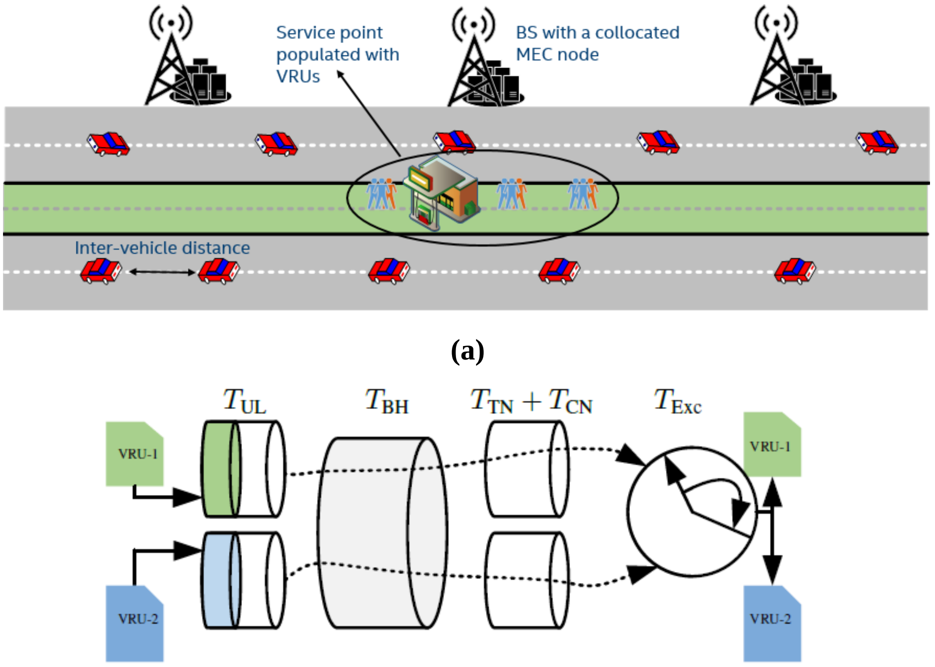
The figure shows that platoons of up to 30 or even 50 members can be supported if the RRC Idle or the RRC Inactive states are used by the V2X devices, respectively. For larger platoons, it is required that the platoon leader remains in the RRC Connected state, otherwise the E2E latency requirement cannot be met. Finally, the E2E delay requirement cannot be met for platoons with 60 or more members regardless of the RRC state of the platoon leader.

In the remainder of the section, we concentrate on communications via the Uu-interface, focusing on the VRU use case, where the transmitting entity is a VRU and the recipients are clusters of vehicles. We aim to decrease the E2E latency and we propose a new network architecture incorporating MEC infrastructure, co-deployed with the RAN. The ultimate goal of the proposed network architecture variant is to illustrate the E2E latency gains that can be achieved, when a VRU packet does not traverse the whole, “distant cloud”-based core network (CN). It should be noted, that, in contrast to the previous performance assessment, the UEs are assumed always RRC connected.

4.2.4.2 MEC-assisted End-to-End Latency Evaluations for C-V2X Communications

Focusing on the cellular V2X (C-V2X) technology, the architecture of the cellular network is expected to have a vital impact on the support of delay-intolerant V2X services. This occurs because the E2E latency of C-V2X signalling is limited by the quality and dimensioning of the cellular infrastructure, i.e., the capacity of backhaul connections, as well as the delays introduced by both the core network (CN), and the transport network (TN). In this section, we argue that stringent latency requirements posed by the V2X system can be satisfied by introducing multi-access edge computing (MEC) technology to the cellular network architecture. Leveraging its ability to provide processing capabilities at the cellular network’s edge, an overlaid MEC deployment is expected to assist vehicles in achieving low packet delays, due to its close proximity to the end users. As a consequence, concentrating on the VRU use case, which studies the safe interaction between vehicles and non-vehicle road users (pedestrians, motorbikes, etc) [5GAA17] via the exchange of periodic cooperative awareness messages (CAM), we aim to reveal the latency-related benefits by introducing MEC system deployment over a state-of-the-art cellular network. Through simulations, we provide clear evidence that the deployment of MEC infrastructure can substantially reduce the E2E communication latency, assuming realistic system parameter values. Our study exploits the existing cellular infrastructure and assumes V2X communication via the Uu radio interface [3GPP36.305].

Throughout the section, a freeway road environment is assumed, consisting of one lane per direction, where a VRU is assumed to interact with vehicles and, possibly, other users on the road. The setup is graphically illustrated in Figure 4-24a. A straightforward example is the one of safety-related applications, in which periodically generated VRU messages can be exploited for crash prevention purposes. Regarding the conventional cellular network architecture approach, the one-way CAM messaging latency components are shown in Figure 4-22b. Moreover, to account for the nature of CAM messages, where the E2E latency is dependent on the successful reception of the packets by the destined vehicles, we resort to the concept of *location-based vehicle clustering*. According to this approach and, based on location availability, each VRU defines a cluster of close-by vehicles and a *cluster-based multicast transmission* takes place in the DL.


Service point populated with VRUs
BS with a collocated MEC node
Inter-vehicle distance
T_{UL}
T_{BH}
$T_{\text{TN}} + T_{\text{CN}}$
T_{Exc}
VRU-1
VRU-2
VRU-1
VRU-2

(a)

(b)

Figure 4-22 a) The investigated two-lane freeway scenario consisting of a cluster of VRUs, vehicles and eNBs with MEC servers co-located to the radio nodes b) One-way signalling latency for two VRUs - conventional approach.

Always focusing on the state-of-the-art network architecture approach, the E2E latency is expressed as $T_{E2E} = T_{UL} + 2(T_{BH} + T_{TN} + T_{CN}) + T_{Exc} + T_{DL}$, where T_{UL}, T_{DL} and T_{Exc} represent the uplink, downlink and execution latencies, respectively, whereas the latencies introduced by the Backhaul (BH), Transport Network (TN) and Core Network (CN) are termed after as T_{BH}, T_{TN} and T_{CN}, respectively. For a network architecture exploiting MEC resources, the latter three latency components are eliminated, thus, reducing the overall experienced E2E latency. Detailed information on the modelling of the various E2E latency components can be found in [EFS18].

From a performance standpoint, Figure 4-23 shows the average E2E signalling latency with and without MEC host deployment, both as a whole and component-wise, as a function of the number of VRUs. Clearly, focusing on a given number of VRUs, MEC utilization provides a lower E2E latency (the observed gains are in the range of 66%-80%), due to the exploitation of processing resource proximity offered by the MEC host. Additionally, we observe an increasing behaviour of the latency along with the VRU density, which is due to the increasing demand of the available resources. First, for the radio transmission latency components, as the number of VRUs increases, the available resources per VRU decrease due to the adopted equal resource allocation scheme. Similar explanations hold for the BH and the execution latencies. The TN and CN latencies were modelled as uniformly distributed random variables based on the recent measurement campaign reported in [SNH+18].

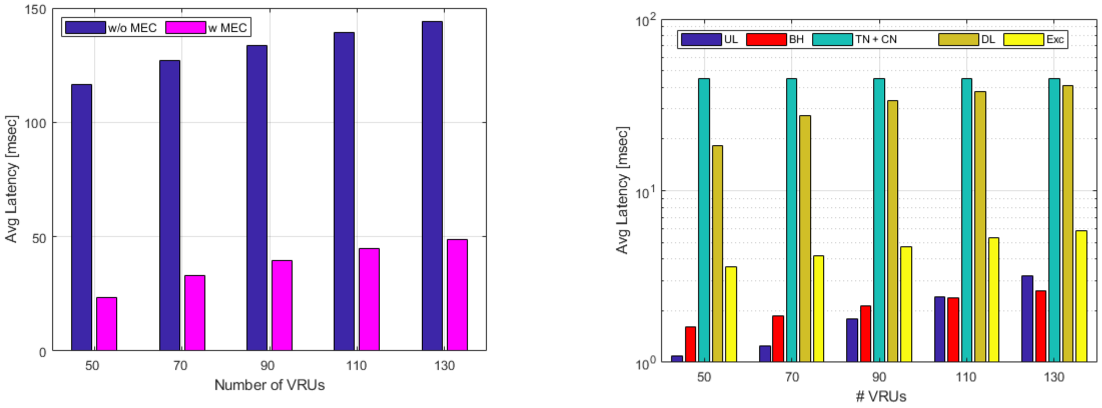


Figure 4-23 (Left) Average E2E latency as a function of the number of VRUs for 30 vehicles per unit length of the road. (Right) Component-wise latency breakdown.

As a next step, we investigated different simulation scenarios by varying the values of other main system parameters (e.g. the vehicles density, which is 30 vehicles for the reported results) as well as the vehicles' cluster size impact on the experienced latency. The experienced average E2E latency when the density of vehicles (i.e., number of vehicles per unit length of the road) increases is shown in the left hand side of Figure . 4-26. The slight reduction the average latency can be explained as follows. As the number of vehicles increases, the distance between the cluster members and the serving gNB decreases, thus the downlink latency decreases. Finally, the right hand side of Figure 4-26 shows the effect of increasing the cluster size. Since our model computes the average downlink latency based on the successful reception of all cluster members, as the size of such cluster increases, the average downlink latency increases as well.

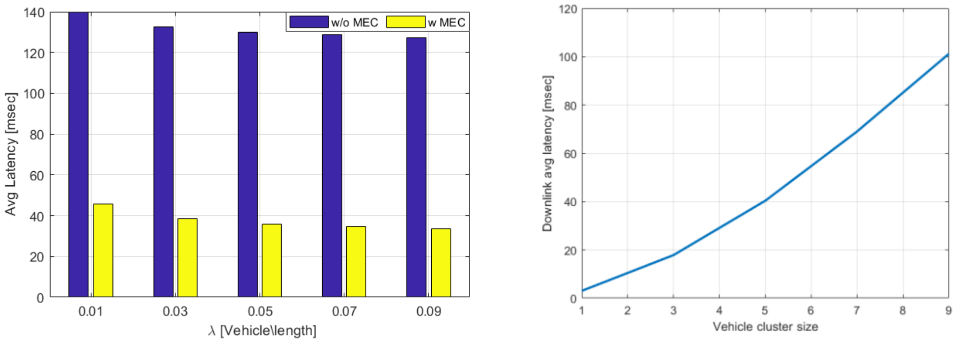


Figure 4-24 (Left) Average E2E latency as a function of the vehicles density (Right) Average downlink latency as a function of the cluster size

As a summary, with the aim of minimizing E2E signalling latency, we have proposed a MEC-assisted network architecture, according to which MEC hosts are collocated with eNBs, thus, they can receive and process VRU messages at the edge of the access network. By means of numerical evaluation, it has been observed that, for some of the investigated system parameterizations, the proposed overlaid deployment of MEC hosts offers up to 80% average gains in latency reduction, as compared to the conventional network architecture.

4.2.5 Utilization of Prediction in Small-Cell Mobility

Enhancing URLLC communications and boosting robustness over 5G wireless networks through predictive network control (PNC) algorithms for MC and centralized forwarding is the innovative approach considered here. DC, a primary aim of MC, will be crucial in supporting tight interworking in HetNets. Thus, this study focuses on URLLC service provision in an MC mobility scenario where macros and small cells are deployed at non-overlapping carrier frequencies. In fact, in this case only small cell mobility events are considered in an established multi-connection. Moreover, to deal with the user plane data flow, a split bearer architecture for MC is examined since it is a promising enabler to cope with cell densification management and MC complexity in the RAN. DC/MC as an URLLC promising enabler, summarized in Table 3-3 where novel functionalities on PDCP PDU duplication are exposed, is studied here from the mobility point of view.

This investigation proposes a centralized predictive flow controller to handle MC for URLLC services. The controller is based on a discrete time model with binary controls as in [SW18b]. The predictive essence of the controller is built upon CSI and nodes buffer state. Even though system level simulations and results are shown for DC as recommended in 5G NR Release 15, the solution is applicable to the general MC case.

We consider a scenario where a UE is simultaneously connected to two different cells, a primary cell (PCell) and a secondary cell (SCell), through a master NR gNB (MgNB) and a secondary NR gNB (SgNB) respectively. MgNB and SgNBs operate in different carrier frequencies as shown in Figure 4-25 (Left). Moreover, data forwarding between the secondary gNBs is not considered; as a result, SgNB management interruption time is not assumed close to zero. MC is only applicable to UEs in RRC connected mode. Hence, a DC-enabled UE has two identities (radio network temporary identifier, RNTI) one C-RNTI in the master cell group and another C-RNTI in the secondary group. Additionally, the user data bearer split in the RAN is considered for the MgNB, so the data from one radio bearer can be either forwarded directly to the UE or relayed to a SgNB via a (wired) Xn interface. as shown in Figure 4-25 (Right). In terms of mobility, Figure 4-25 (Left) highlights the SCell mobility events such as SCell Add A4, SCell Change A6, and SCell Release A2 considered to activate/deactivate the proposed controller.

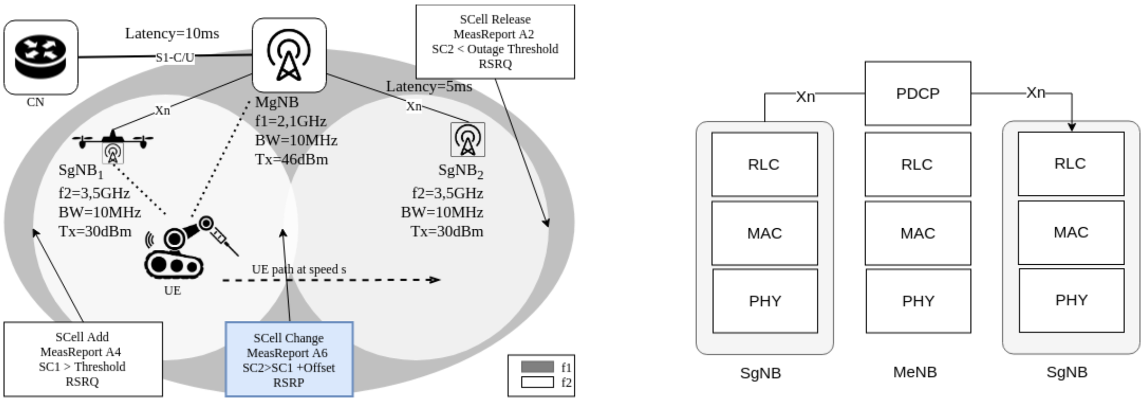


Figure 4-25 (Left) MC Mobility Scenario, (Right) 5G NR Layered Architecture

Figure 4-26 describes a scenario, in which buffer states and UE communication links are time-variant; as a result, optimal path selection must be adaptive. The common forwarding method of PDCP-PDUs is built for handling single connectivity (SC). The proposed PNC algorithm can handle multiple-connections and relies on multiple-step-ahead prediction taking advantage of periodic wideband CQI (wbCQI) reports to improve packet data unit routing, forwarding and duplication during mobility events. The periodic wbCQI reports are assumed to be configured by the upper layer (RLC). The wbCQI reports are available at different instants in time for each wireless link, e.g. *wbCQI$_1$* corresponds to the report sent through link *1* at *t_0* when only SC with the MgNB is available.

The analytical model is based on a controlled queuing system with an exponential inter-arrival rate λ at the MgNB and service rate μ for each gNB, assuming, without loss of generality, equal processing capabilities for all gNBs. The model is depicted in Figure 4-26(Left). The centralized controller defined for a discrete time packed level proposes a queue vector $q_t \in N^{n_q}$, an arrival vector a_t, routing matrix R_t, and a control vector v_t. Data forwarding decisions are obtained as the convex optimization of a utility function, under the probabilistic knowledge of the future behaviour of the channel state based on a known UE trajectory. The system state evolution can be written as [SW18b]:

$$q_{t+1} = q_t + R_t v_t + a_t$$

Where q_t is a vector composed by MgNB buffer state q_t^0, SgNB$_1$ buffer state q_t^1, SgNB$_2$ q_t^2.

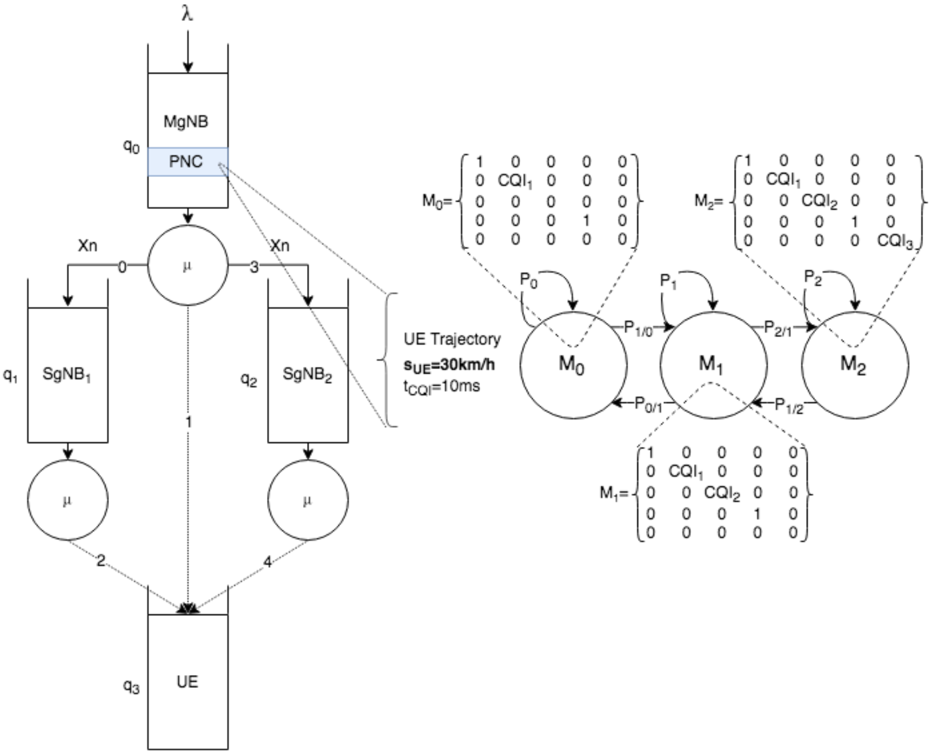


Figure 4-26 (Left) PNC Queuing Model, (Right) Channel State Matrices and Channel State Markov Chain

The channel matrices are diagonal matrices capturing the transmission success probability in the links between the queues as depicted in in Figure 4-26 (Right). For example, an initial channel state matrix M_0 captures wired Xn links $\{0,3\}$ with the highest success probability, $m_{1,1} = 1, m_{4,4} = 1$, respectively. The CQI values in the channel state matrices are proportional to standard wbCQI indexes, assuming a maximum probability for the maximum standard wbCQI=15. In fact, these values were obtained based on system level simulations for a certain UE trajectory at a predefined speed, further information can be found in Annex 7.5. The evolution of the channel is represented by a Markov chain as shown in Figure 4-27. There is a pre-configured Markov chain for a specific UE speed.

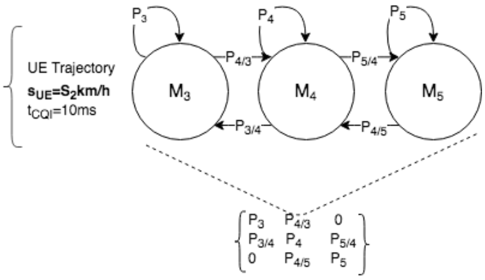


Figure 4-27 Channel State Evolution Markov Chain

The algorithm in the PNC controller is activated by the reception of the MeasReport A6. The required structures to store the received wbCQI, structures for the pre-configured channel state matrices, buffer states and control vectors are instantiated. The latter with the aim to compute the optimization variables $u_t, u_{t+1}, \dots, u_{t+H}$, to finally apply at time t the control value u_t. Hence, controlling the forwarding of packets between the nodes SgNB, or UE. The control value u_t can activate a column in the routing matrix representing the forwarding of packets in SC, DC or MC. The utility function is of quadratic type and contains expectations of future system quantities:

$$\min_{u_t, u_{t+1}, \dots} E\left[\sum_{i=t+1}^{t+H} q_i^T Q q_i + u_{i-1}^T R u_{i-1}\right].$$

For generality purposes, the symmetric weights matrices Q and R have been introduced in order to penalize information buffering and link activation, respectively. Moreover, H is the horizon variable which defines the timesteps considered in the prediction. For the case of system level simulations, a time-step is considered as 0.1ms. The controller calculates an optimal state trajectory for a horizon (H) according to the previous cost function. For the optimal trajectory of the system there corresponds an optimal control trajectory.

System level simulations results show gains in terms of instant throughput in SCell Change A6 event as shown in Figure 4-28. We compare the packet forwarding in the PDCP layer for an autonomous SCell Change based on A6 event with an PNC controlled strategy, in a HetNet serving eMBB users and URLLC users (Annex 7.5). The gains are about 10% of throughput during the execution of a SCell change process. Hence, we recommend a PNC controlled strategy to forward PDCP PDUs in a MC connectivity scenario with a split bearer architecture. The PNC strategy is based on periodic CSI reports captured in the PDCP layer in the MgNB which is acting as the base station for the primary cell (PCell).

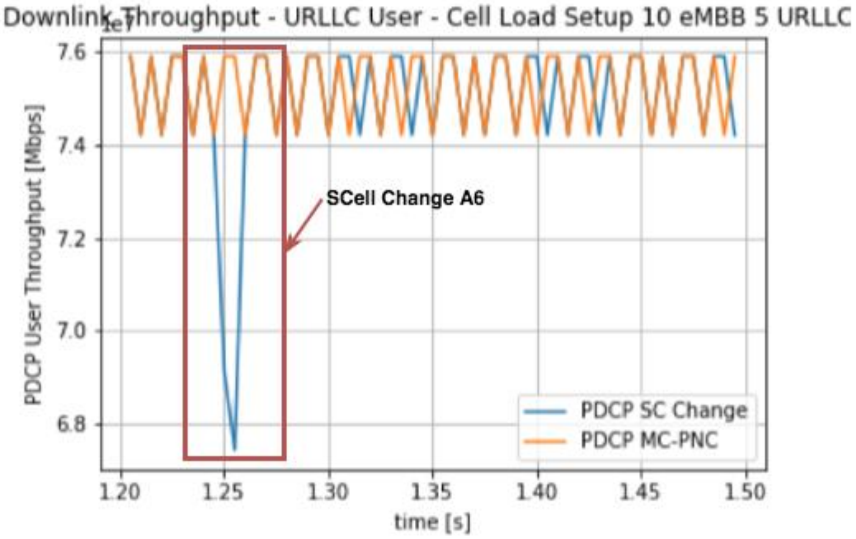


Figure 4-28 PNC Downlink Throughput Performance

4.2.6 Mobility signaling for high-speed trains

The high-speed train scenario poses specific requirements to both PHY and RRM. At PHY level, large and varying Doppler shifts necessitate special care with respect to frequency offset compensation. At RRM level, specifically serving cell management becomes prone to signaling failures due to fast changing large scale propagation (antenna gains and path loss). The default signaling procedures have difficulties to reliably handle many handovers, i.e. the changes of the primary cell (PCell) that carries all RRC signaling. Regularly handovers even take place back-to-back, limited just by the (processing) speed of RRM signaling towards mobiles and between base stations within the NG-RAN.

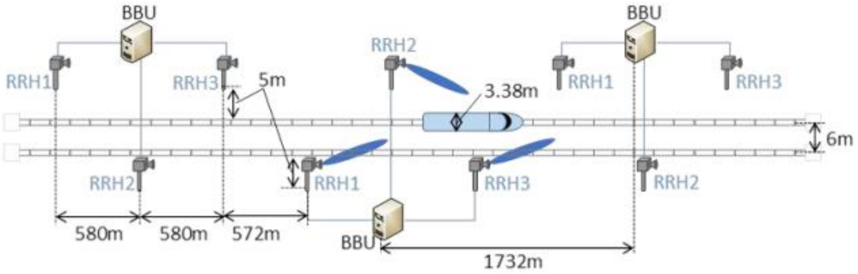


Figure 4-29 3GPP deployment scenario for high-speed trains at 30 GHz (TR 38.913 Figure 6.1.5-2)

These issues can be solved in a straight forward way by placing remote radio heads (RRH) along the track such that their radiation patterns are aligned uni-directionally along only one direction of the track as shown in Figure 4-29. Combined with a single frequency network (SFN), i.e. all RRH transmitting the same signal, Doppler shift becomes mostly constant and handovers become seldom, occurring only at boundaries between two RRH clusters forming an SFN. The drawback of uni-directional antenna alignment is that it utilizes only half the available spatial dimensions: a train transmits and receives only into either forward or backward direction, depending on its direction of travel, but not in both. On top, SFN trades spatial reuse for reduced handovers, removing the possibility to transmit different user data over different RRH, from which specifically long trains with two or more relay antenna positions on their roofs could benefit. These drawbacks motivate the following RRM signaling improvements as an alternative solution to the high-speed train scenario with the prospect to more than double the achievable throughput while retaining a comparable level of robustness.

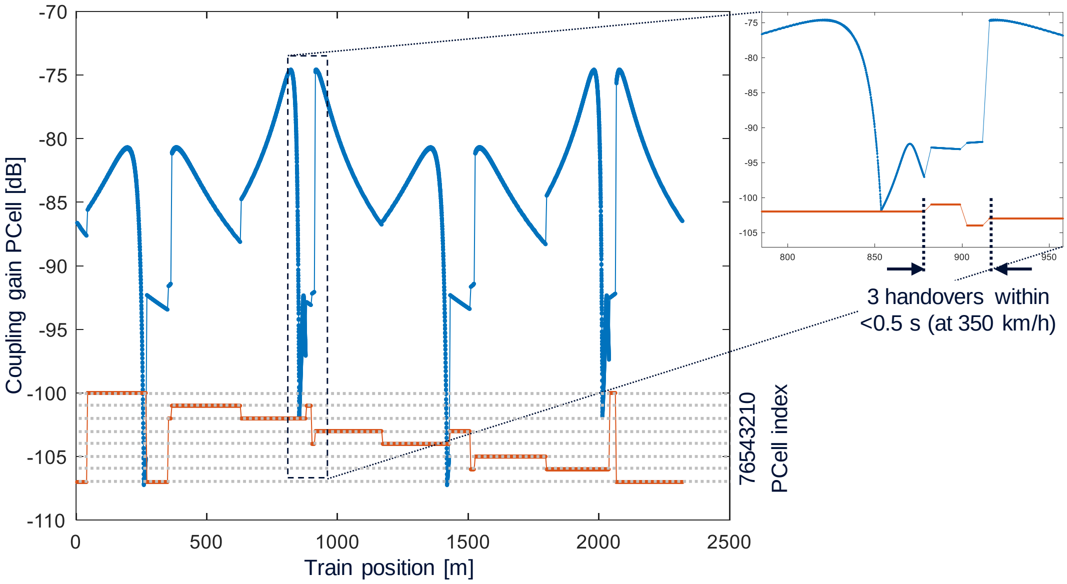


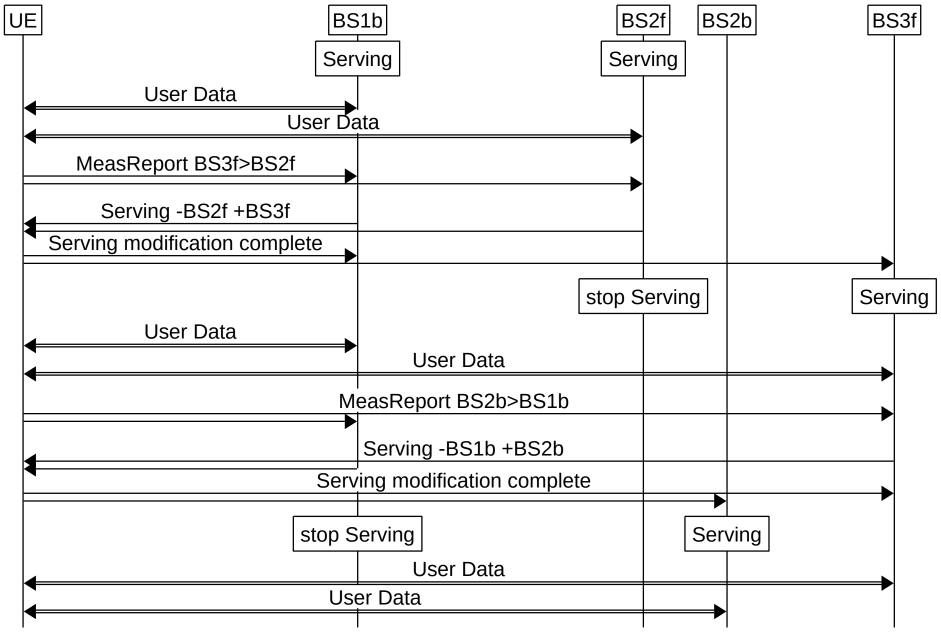
Figure 4-30 Coupling loss and according handovers with standard signaling simulated for part of the scenario in Figure 4-29 but with bi-directional RRH antennas (two cells per RRH position)

To exemplify the behavior of standard Rel-15 handover signaling in case every RRH location in Figure 4-29 would be equipped with a second RRH facing left along the opposite direction of the track, Figure 4-30 shows the coupling loss and according handovers a train experiences that travels at 350 km/h from left to right past the three center RRH positions (with the blue beams) plus RRH1 on the right. The cut-out on the top right shows that the three handovers take place within less than half a second (less than about 50 m traveling distance) when the train passes by the center RRH2 position.

Handover robustness can be improved and probability of handover failures and prolonged interruption times after a handover failure can be reduced by employing

- Multi-connectivity (utilizing PDCP split bearer with packet distribution for user plane), i.e. the train connects to the two cells (=RRHs) with the strongest receive signal, combined with
- Split SRB (signaling radio bearer) on these two cells, i.e. RRC messages are carried over a PDCP split bearer with packet duplication enabled.

This still requires the (contention-free) random access procedure to be carried out each time, which (shortly) interrupts data transfer, but more importantly, which might fail, causing the UE to revert to back to the contention-based random access as part of re-establishment procedure. The RA-less handover (cf. section 4.2.1) can improve handover robustness but it does not reduce the number of handovers. With dual-connectivity applied to the SRB, all serving cells always carry all RRC messages and we propose not to distinguish a PCell but consider all cells equal with respect to signaling. The resulting handover, or more precisely serving cell management, signaling is shown in Figure 4-31. BS2f and BS2b are the two cells located at position of the center RRC2 of Figure 4-29, BS2f facing left and BS2b facing right. Similar for BS1b and BS3f, which are located at position of RRC1 facing right and RRC3 facing left, respectively. As shown, the three back-to-back handovers (cf. Figure 4-30) are replaced by two serving cell additions and removals. Since one serving cell is always available while the other one is changed, the SRB continues to use HARQ and RLC, providing retransmissions to avoid lost signaling messages.



http://msc-generator.sourceforge.net v6.3.7

Figure 4-31 Proposed RRC signaling for serving cell management with SRB duplication and NG-RAN internal primary cell and cell group management, i.e. without UE participation

For dual connectivity within the same carrier frequency, the train (i.e. UE) must be able to spatially separate its front and back link. For this, the train can either employ an active antenna array or, as done here, a pair of directional antennas, one pointing into the train's direction of travel and the other into the opposite direction. The resulting two coupling gains and two serving cell indices are shown in Figure 4-32. The higher coupling gain in comparison to Figure 4-30 is due to the higher antenna gain of the directional train antenna vs. the omnidirectional antenna used in case of single connectivity. The absolute values do not matter, though, because the system is interference-limited. As can be seen, the front link suffers a fast drop in coupling gain within the about 200 ms from measurement report until the cell change has been completed. This renders the source cell between 15 dB and 50 dB weaker than the target cell right before switchover. Furthermore, the strongest of the two serving cells changes back and forth between the two cell, which normally causes a PCell change, i.e. a handover. For example, cell 0 becomes stronger then cell 7 at position 10 m and vice versa cell 7 again stronger than cell 0 at 210 m. Then cell 0 is replaced by cell 2 and cell 7 by cell 1 and the same repeats with cells 1 and 2 (cell 2 becomes stronger than cell 1 and vice versa), and so on. For standard NR Rel-15 handover signaling[3], we therefore assume that the network applies SON (self-organizing network) mechanisms to always force the PCell on the back link. This avoids both unnecessary handovers between PCell and SCell every time the SCell becomes stronger than the PCell and it avoids potential handover failures if signaling messages would be carried over the front link during its steep dip in coupling gain.

For the proposed novel signaling, each time another cell becomes stronger than either of the two current serving cells, a cell serving cell modification procedure is triggered as explained before in Figure 4-31. Reliability is provided through the respectively unchanged cell. No further (SON) mechanisms are necessary beyond that. For both dual connectivity approaches, standard handover signaling with SON and the proposed novel serving cell management, the resulting current PCell and SCell respectively the two serving cells are the same and are shown in the lower part Figure 4-32.

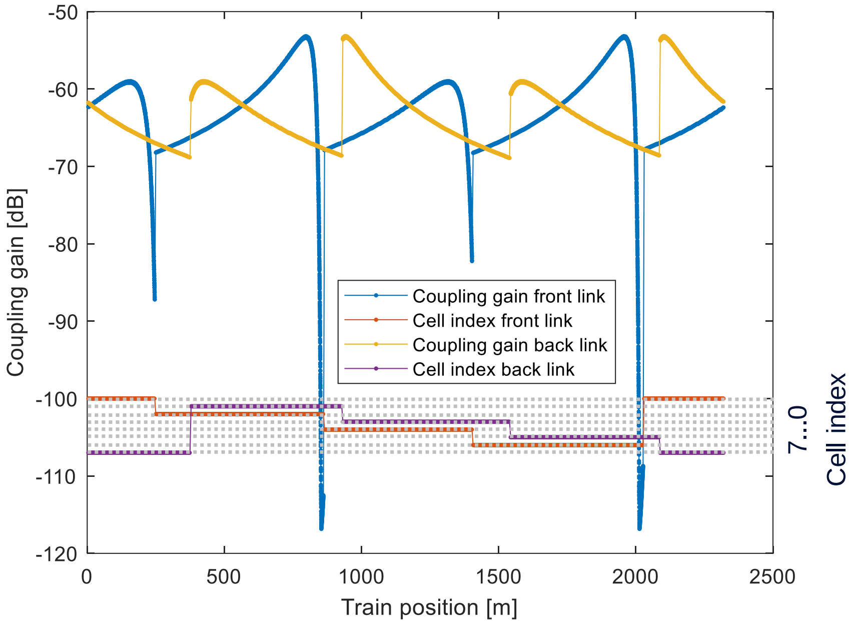


Figure 4-32 Coupling loss and according serving cell changes for part of the scenario in Figure 4-29 but with bi-directional RRH and bi-directional train antennas (two cells per RRH position, two antennas per train)

7 lists the durations that zero, one or two cells are available for carrying user data, i.e. have at least one non-suspended data radio bearer (DRB), along with the resulting percentages for outage (zero cells), degraded performance (1 out of 2 cells), and overall availability (at least one cell), respectively. During the time the train travels one inter-TRP distance (580 m), for single connectivity, 3 handovers, for dual connectivity 1 handover (PCell change, i.e. zero cells

[3] NR Rel-15 does not specify dual connectivity within the same carrier, though support could be added without changes to RRC signaling.

available) and 1 SCell change (1 cell available, namely the PCell), and with the proposed signaling 2 cell changes (1 cell available, namely the respective other serving cell) take place.

Table 4-7 Comparison of minimum outage times of single connectivity, and dual connectivity with SON and dual connectivity with proposed signalling (best case without signalling failures)

	Availability at 350 km/h train speed, averaged over one inter-TRP distance of 580 m respectively over 23862 slots of 250 ms						
Cells available for user data transmission	0 cells		1 cell		2 cells		Overall (>0 cells)
Single Frequency Network (12 cell cluster)	47 slots	0.2 %	23815 slots	99.8 %	n. a.	n. a.	99.8 %
Single connectivity	561 slots	2.4 %	23300 slots	97.6 %	n. a.	n. a.	97.6 %
Dual connectivity (SON with standard signaling)	163 slots	0.7 %	167 slots	0.7 %	23532 slots	98.6 %	99.3 %
Dual connectivity (proposed signaling)	0 slots	0 %	330 slots	1.4 %	23532 slots	98.6 %	100 %

For comparison, single connectivity and SFN performance are given. The SFN provides 100 % availability within each SFN cell cluster. Between two adjacent clusters, handover performance can be assumed to be comparable to single connectivity. Accordingly, for a SFN cluster size of e.g. 12 cells, the average outage is a 1/12 of that of single connectivity. As can be seen, dual connectivity with standard NR Rel-15 signaling already achieves a high availability of 99.3 % in the absence of signaling failures. At least in the evaluated scenario, through suitable selection and orientation of antennas, signaling failures can be avoided.

The proposed new signaling scheme reduces outage to practically zero, enabling URLLC services. A signaling failure increases the duration of degraded (i.e. 1 cell available) performance by the time needed for retransmitting the lost RRC message but does not lead to an outage (zero cells available). Both dual connectivity solutions provide about twice the throughput of SFN and single connectivity (as already noted, the system is interference-limited; accordingly, the best transport format is used almost all the time except during cell change).

Although not part of the presented study, it should be noted that a further throughput improvement may be achievable by adding a second antenna location to the train resulting in two pairs of directional antennas (or two active antenna arrays) at either end of the train. SFN cannot benefit from additional train antennas, but alternatively two SFNs, one set of RRH antennas pointing into the train's direction of travel, the other set of RRH antennas into the opposite direction, may be paired to double the train's throughput (reaching that of dual connectivity with one antenna position per train). Handovers between SFN clusters are then carried out as in case of dual connectivity.

4.3 Connectivity optimizations for device-to-device communications and relaying paths

This section investigates D2D communications from different perspectives. With the exponentially increasing number of connected devices and a rising demand of higher data rate applications D2D communications can be seen as an efficient technique to offload the traffic, extend the coverage and reduce the interference. In addition, D2D communication is of particular interest to optimize the radio resource allocation as well as the energy consumption in the case of mMTC applications. In this section, different aspects related to D2D communications are investigated.

The first investigated problem is a resource optimization problem in a heterogeneous network composed of both D2D and non-D2D users. The carried out study takes into account the dynamic traffic arrivals of the users, their average delay requirements as well as their energy consumption. Since a centralized resource allocation policy requires high signalling overhead, the focus of this work is to develop a fully distributed scheduling policy in which each user or device decides locally to transmit or not without requiring information exchanges between the transmitters. The developed policy is characterized by being both throughput optimal and energy efficient.

Another aspect studied in this section concerns the problem of power minimization in a dense MTC network. A strategy is proposed consisting of a simple threshold based policy that can be implemented in a distributed way in the network.

The third investigated aspect concerns D2D relaying for eMBB services. A distributed relay selection strategy that takes into consideration the time varying nature of the traffic and the signaling overhead in the network (limited feedback) is introduced. The last aspect concerns D2D relaying schemes for mMTC services. A complete D2D relaying scheme is proposed (discovery protocol and transmission phase) for the scenarios where the device is in coverage or at the edge of uplink coverage. For the transmission phase the interest of using a Chase-Combining HARQ scheme is studied. The section ends up by introducing a D2D discovery protocol inspired by the 802.11 RTS-CTS protocol for optimizing energy consumption.

4.3.1 Resource optimization for eMBB and mMTC

4.3.1.1 Stochastic resource optimization for heterogeneous architecture

We investigate the problem of resource optimization in a network composed of both D2D and non-D2D users configuration (heterogeneous architecture) that takes into account the dynamic traffic arrivals of the users, their average delay requirements as well as their energy consumption. In particular, we are interested in this section in user scheduling with the objective to develop a distributed optimal policy, under the assumption that each transmitter knows only its own Channel State Information (CSI) (i.e. the CSI within its own receiver). The objective of the work is to construct a distributed scheduling scheme that takes into account the dynamic traffic arrival of the users and which is characterized by being both throughput optimal and energy efficient. A scheduling scheme is said to be throughput optimal if it can support any feasible incoming rate, i.e. it stabilizes all the queues within the network whenever it is possible to do so. The previous literature on this matter [JW10] solely focused on the throughput optimality aspect of the scheduling schemes and overlooked the energy consumption side. To incorporate the energy consumption aspect, our proposed scheme involves giving each link the freedom to transition between AWAKE and SLEEP states. However, unlike the conventional duty cycling that has been previously proposed in the literature, the sleeping duration of each link is not fixed and is calibrated with the aim of optimizing a particular objective function. In fact, we show that by jointly controlling both the sleeping and back-off duration of each link with the aim of optimizing a certain objective function, the power consumption of each link can be reduced while still maintaining the ability to withstand any feasible arrival rate.

The proposed scheme in our work belongs to the family of Carrier Sense Multiple Access (CSMA) Medium Access Control (MAC) protocols where transmitters in the network listen to the medium before proceeding to the transmission phase. More specifically, an active transmitter (i.e. not asleep) waits for a certain duration before transmitting, called the back-off time. While waiting, it keeps sensing the environment to spot any conflicting transmissions. If an interfering transmission is spotted, the transmitter stops immediately its back-off timer and waits for the medium to be free to resume it. Motivated by reducing the power consumption, we provide each transmitter the freedom to transition between two states:

- SLEEP state: in this state, power consumption is minor and no sensing of the environment takes place (the transmitter's radio is OFF);
- AWAKE state: the transmitter's radio is ON and the CSMA scheme takes place.

The decision to either wake-up/sleep is dictated by an appropriate timer. When a transmitter decides to sleep, it picks an exponentially distributed wake-up time after which it wakes up. Once the transmitter is awake, it picks an exponentially chosen sleep timer after which it goes back to sleep. Our goal becomes to properly calibrate the means of both, the sleeping and back-off timers in a certain way to guarantee throughput optimality [MAE18] with low energy consumption. An example of our proposed scheme with two transmitter-receiver links is depicted in Figure 4-33.

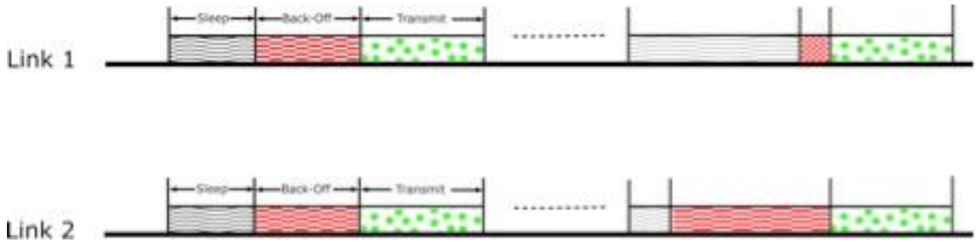


Figure 4-33: Example of proposed CSMA scheme

We formulated an optimization problem with the aforementioned means as control variables. We were able to show that the optimization problem can be solved in a distributed manner where each node in the network simply monitors its past service rate and awake duration. At the optimal point, we showed that the throughput requirement of each node is satisfied and nodes wake-up just as needed. More precisely, a new parameter is introduced and is assigned to each node based on the desired power-delay trade-off. For any chosen feasible parameter, we were able to theoretically prove that the proposed scheme remains throughput optimal [MAE18].

The interest of this parameter comes from the fact that 5G applications present a mixture of both delay-sensitive and delay-tolerant applications, which can be addressed by simply calibrating the parameter assignment.

We consider a network of low power RF transceivers proposed by the industry where power consumption is detailed in [Texas13]. More specifically, a heterogeneous case whith 3 groups of service flows is considered, where each service flow is transmitted over a transmitter-receiver link. We consider also 4 links (i.e. 4 service flows) in each group:

- Group 1: This group is made of links that are delay sensitive but can tolerate a high power consumption;
- Group 2: This group is made of links that fall between the two extremes, they require a moderate power consumption without introducing a lot of delay;
- Group 3 This group is made of links that can tolerate long delays however they are extremely power limited.

We assume that the average arrival rate for each link is 0.077 packet/slot and we assign the power-delay trade-off according to each group’s requirement. We compare the performance of our proposed scheme to the throughput optimal adaptive CSMA [JW10]. In Figure 4-34, one can clearly see how our proposed scheme is able to satisfy the throughput requirements of each node of the groups while providing huge power gains (3 to 5 times less power consumption). . More details on the proposed solution and the obtained results can be found in [MAE18]. The main recommendation is that by adapting CSMA (as it is done in this section), multiple delay requirements can be met (i.e multiple services) and in parallel a huge power gain can be achieved. This implies that the adaptive CSMA can be used jointly for eMBB and MTC (with low power consumption) for small to moderate density of devices..

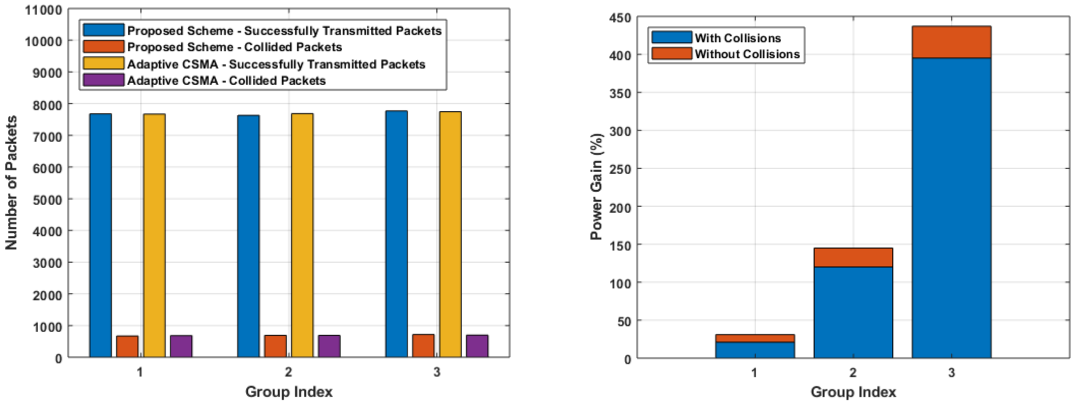


Figure 4-34 (Left) Average throughput achieved by adaptive CSMA and our proposed scheme. (Right) average power gains by our scheme with respect to the adaptive CSMA [JW10] with and without collisions

4.3.1.2 Power consumption reduction for mMTC

The high density and stochastic activity of mMTC requires the development of a low complex and distributed transmission policy that requires a very low signaling overhead, while at the same time, minimizes the total power consumed by the MTC devices. In this section, we address this issue, i.e. the problem of power minimization in a dense MTC network, and develop a simple threshold based policy that can be implemented in a distributed way in the network. The proposed solution takes into consideration the dynamic activity of the MTC and wireless channel conditions and requires very low signaling overhead in the network. This is an advantage as compared to most of existing works in this area (i.e. power control in dense networks) that focuses solely on the wireless links, i.e. CSI, without taking into account the traffic patterns and/or the queues of the users. In fact, the CSI reveals the instantaneous transmission opportunities at the physical layer and the queue state information (QSI) reveals the urgency of the data flows. A good transmission policy must take into account both the CSI and the QSI.

In more details, we consider a network composed of a high number of MTC devices that have stochastic traffic (activity) to transmit to an access point (or gNB). The MTC devices may interfere with each other if they transmit at the same time and we consider that their signals are decoded correctly if the received Signal to Interference and Noise Ratio (SINR) is higher than a predefined threshold. In this work, instead of considering game theory or using heuristic algorithms, we approach the problem in a different way. First, we formulate the overall power control problem as a stochastic optimization problem (due to fast fading and stochastic traffic arrivals) and our objective is to find the global optimal (or near optimal) solution in the considered dense network. We then overcome the dimension problem by using the mean field approach. This consists of neglecting the behavior of individual users by only considering the proportion of users in a certain state, which allows us to move from a stochastic optimization problem to a continuous-time deterministic one. We formulate the optimal control problem based on this deterministic approach and solve it by characterizing a solution of the Hamilton-Jacobi-Bellman (HJB) equation. The main challenges in this case are that the equations are fully coupled, which makes the solution of the HJB very complex. We handle these challenges by characterizing the optimal equilibrium point of the dynamic system with respect to the control variable. Then, we prove that the cost function is convex (in all possible equilibrium points) and derive the optimal policy (i.e. that satisfies the HJB equation). This optimal policy (of the asymptotic or mean field regime) turns out to a simple threshold type of policy that can be easily applied in the original stochastic system (and in a distributed way) and provides nearly optimal performance. More details on these results can be found in [LAD18].

The aforementioned theoretical results are also corroborated by numerical ones. We provide in Table 4-8 an example of such results where we compare the performance of our proposed mean field solution with respect to the numerical optimal solution of the original stochastic problem obtained through value iteration (VI) [LAD18]. We consider a network of 10 transmitters in the system where each user has at most one packet to transmit. This is motivated by MTC services, e.g. sensors (that measure temperature), where each machine or sensor has few packets to transmit. The use of only 10 transmitters is due to the fact that the solution of VI is out of reach for high numbers of devices (due to the complexity of VI) and the comparison between our scheme and VI may not be feasible. The setting simulated here is also of interest since we observe in Table 4-8 that our proposed solution is nearly optimal across all values of traffic arrival rate ρ (packet/slot). In fact, the relative error between mean field and VI (defined as $\frac{P_{MF}-P_{VI}}{P_{VI}} \times 100$) is very small, where P_{VI} and P_{MF} are respectively the average consumed powers by VI and mean field policies. This suggests that mean field approach, that highly simplifies the stochastic resource optimization framework, can be used in networks with moderate numbers of users. The main conclusion is that a binary power control, which is simple to implement in practice with low signaling overhead, can achieve a tradeoff between energy consumption and average delay in the scenario where the traffic arrival is low and for Gilbert-Eliot channel model [LAD18].

Table 4-8 Comparison between our mean-field policy and Value Iteration (VI)

ρ	0.05	0.1	0.2	0.3
Relative Error (%)	0.1233	0.1158	0.164	0.3494

In order to assess the performance in a more realistic context, we consider 100 users in the system and we compare our policy with two other policies. In the first one, the devices transmit always at max power whenever they have packets and in the second one the power is allocated using the successive convex approximation (SCA) approach [ASP14]. Figure 4-35 depicts the performance in terms of outage probability versus the traffic arrival rate. Outage occurs whenever a packet is not delivered. The performance of our proposed scheme, in terms of outage probability, is depicted in Figure 4-35. This figure demonstrates that our proposed policy outperforms both the maximum power allocation and the SCA-based scheme. In fact, there is a gain of more than 100% between the proposed scheme and the two other existing schemes. To conclude, a simple threshold-based power control, can efficiently achieve a trade-off between minimizing power consumption and average delay in mMTC scenarios.

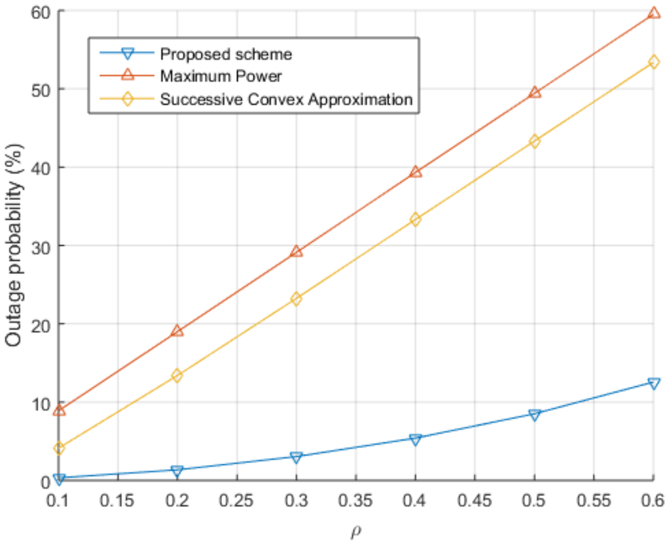


Figure 4-35 Outage probability versus traffic arrival rate ρ

4.3.2 D2D relaying for eMBB and mMTC

4.3.2.1 D2D relaying for eMBB: relay selection and feedback strategy

This section considers the use of D2D communication as a cooperative scheme in order to enhance the downlink communication in a wireless network with heterogeneous architecture. The main goal is to take into account the traffic patterns of the users and the wireless channels to improve the network performance. This can be achieved by selecting the relays that have the appropriate radio conditions to improve the throughput of the users. However, this requires the knowledge of the CSI of all the links in the network (i.e. BS-relay, BS-user, relay-user). Since the CSI feedback is not for free, only a limited number of CSIs can be exchanged in the network. Therefore, the relay selection must depend on the traffic pattern of the users, the radio conditions and feedback capacity of the network. Furthermore, we consider eMBB services for which the QoS is characterized by the queuing stability of the users. The stability of the queues ensures that the packets are served with finite/limited delays. The use of the queuing stability metric is of interest, as it allows comparing the performance of our proposed solutions with the fundamental limit of the system (i.e. achieved by a genie scheme with full knowledge of all network states :all CSIs, queues, etc.),. In addition, we assume that the users and the relay can estimate perfectly their corresponding channel fading coefficients, by decoding the training sequences transmitted on the downlink by the transmitter and the relay. However, we suppose that the users have imperfect knowledge of the fading coefficients of the link between the transmitter and the relay.

In this work, we explore the problem of relay selection and CSI feedback and propose two solutions. In the first one, we consider that the relay selection is done at the BS and prove that the proposed solution is optimal under the assumption of limited feedback. We provide a theoretical evaluation of the bound between the optimal centralized solution under limited feedback and the one with full CSI knowledge. We then develop a distributed relay selection and feedback allocation method in which the decision is made at the user side. The users can then take advantage of their local CSIs knowledge to improve the system throughput. The performance of this policy is compared theoretically and by simulation with that of the ideal system where a genie scheduler with full knowledge of the network states (all CSIs, queues, etc.) selects the relays and schedules the users for transmision. The important outcome of this work is that the proposed distributed solution outperforms the centralized one under the limited feedback assumption. In other words, under the considered limited feedback scenario, one can improve the queueing stability of the network if the decision of relay selection and CSI feedback scheduling is done at the user side (as compared to the case where the decision is done at the BS side). One can refer to [DAD+18] for more details on the aforementioned analysis and obtained results.

We provide in Figure 4-36 an example of the obtained results, where we plot the evolution of the total average queue length with respect to the average arrival rate of the traffic which is assumed to follow a Poisson distribution. We recall that when the queues increase without bound the system becomes unstable and the delay experienced by the users becomes infinite. Therefore, the stability region is characterized by the average arrival rates under which the queues stay bounded. Note that the SNR thresholds for the users and the link BS-relay are respecteively denoted by τ and δ. In Figure 4-36, we compare between the ideal system (genie scheduler with full network states), the optimal centralized solution under limited feedback (denoted by MCR), another suboptimal centralized solution with lower complexity (denoted by CR) and the proposed distributed policy (denoted by RW). One can see that, in terms of average queue length and hence stability region, the distributed solution outperforms the centralized ones. This is mainly due to the fact that, in limited feedback context, the users are aware of their instantaneous CSIs while the BS cannot know the CSI unless it is fed back from the users. Therefore, the users (if the feedback decision is made by the users) can decide to feed back only the good CSI and select the relay with which he has good channel conditions. The main recommendation is that the relay selection must be done at the user side (not at the gNB) as this will achieve a better queueing stability region of the network.

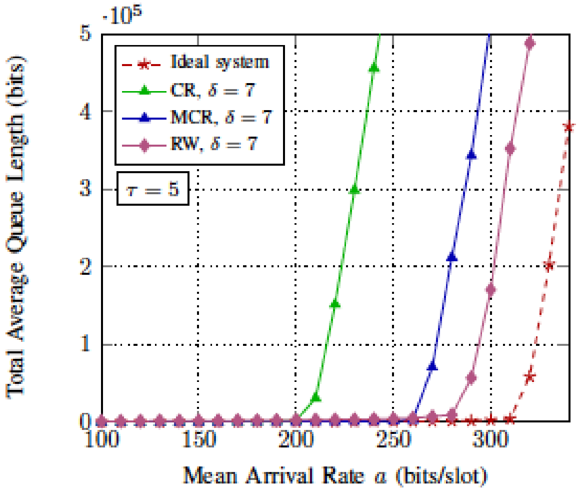


Figure 4-36 Total average queue length vs. mean arrival rate

4.3.2.2 Modeling MTD energy consumption for D2D communications with CC-HARQ scheme

In D2D communications, there are two modes for resource allocation [3GPP36.843]. In mode 1 or Scheduled mode, the base station (BS) schedules and assigns resources. All machine type devices (MTDs) inside the cell then use dedicated resources to transmit as in traditional cellular communications, avoiding intra-cell interference but not inter-cell interference. In mode 2 or Autonomous mode, MTDs autonomously select resources from pools that are preconfigured or given by the base station if they are in coverage. In this mode, depending on the traffic expectations, the mobile operators dimension the resource pools. In Autonomous mode, two or more devices could select the same resource causing interference between them. In our study we focus our analysis on the Autonomous mode, leaving the Scheduled mode for future work.

Retransmission schemes increase the transmission reliability exploiting the temporal diversity of the channel. The most common schemes are automatic repeat request (ARQ) and hybrid ARQ (HARQ) that combines ARQ and forward error correction (FEC) mechanisms. ARQ is the simplest retransmission scheme, in which the same packet is retransmitted if requested by the receiver. In HARQ with Chase combining (CC-HARQ) [CHASE85], the same operations are done as for ARQ at the transmitter side. Nevertheless, the operations are slightly more complex at the receiver side. In our study, the MTDs are the transmitters, while the relays (UEs) are the receivers. ARQ and CC-HARQ are suitable for mMTC applications, since they allow improving the transmission reliability without increasing the complexity of the MTDs.

We propose therefore to use retransmission schemes and more particularly ARQ and CC-HARQ and to compare the performance of the two mechanisms in terms of energy consumption, considering the D2D autonomous resource allocation scheme. We are considering only the transmission of messages from the MTDs to the network since mMTC traffic is usually uplink-dominated. In our analysis, we take into account the impact of Rayleigh fading, log-normal shadowing, and interference caused by other MTDs.

For the network model, the locations of UEs, as well as active MTDs (only MTDs transmitting data at a given moment) form two independent homogeneous Poisson point processes (PPP) with given densities. An MTD transmits its reports using a UE as a relay via a D2D link. D2D links and cellular links are orthogonal. Hence, there is no interference between these systems. An active MTD shares resources with other ones that use the same sub-channel. Hence, a UE acting as a relay suffers interference from other MTD-UE links.

Analytical expressions for the success probability, the average number of transmissions, and the MTD energy consumption are then derived using a stochastic geometry approach. Details on how the analytical expressions are derived can be found in [AVS+18]. The MTD energy consumption is expressed as the sum of the energy consumption during the transmission phase, the reception phase and the idle state:

$$E_{m,1} = P_{m,T}t_{m,T} + P_{m,R}t_{m,R} + P_{m,I}t_{m,I}$$

where $P_{m,T}$, $P_{m,R}$ and $P_{m,I}$ are the MTD power consumption levels (energy consumption per unit time) in Tx state, Rx state, and Idle state respectively; $t_{m,T}$, $t_{m,R}$ and $t_{m,I}$ are the durations of Tx state, Rx state, and Idle state respectively (see Figure 4-37).

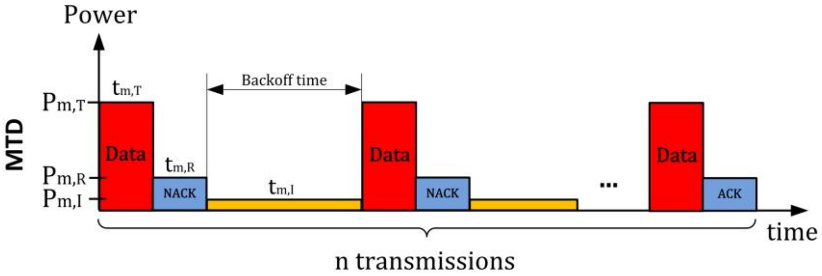


Figure 4-37 Retransmission scheme

We assume that an MTD transmits at a fixed power (no power control) and at a fixed data rate (no link adaptation) due to its low complexity and low cost. Then, we have $t_{m,T} = \frac{L}{R_m}$ where L is the packet size and R_m is the MTD bit rate. In the study, we consider $t_{m,I} = \eta t_{m,T}$ where η is a random integer following an exponential distribution with the mean $\bar{\eta} = 5$. It should be noted that the back-off time does not increase significantly the energy consumption since the MTD stays in idle state during this period of time. We assume that $t_{m,R} = t_{m,T}$ thereby the ACKs will have enough redundancy in order to ensure that they will be received correctly by the MTD.

Then, the average MTD energy consumption in a single transmission can be expressed as:

$$\bar{E}_{m,1} = (P_{m,T} + P_{m,R} + \bar{\eta}P_{m,I})\frac{L}{R_m}$$

And the average of the global energy consumed by the MTD $\bar{E}_{m,G}$ can be derived as:

$$\bar{E}_{m,G}(r) = \bar{E}_{m,1}\bar{T}$$

where $\bar{T}$ denotes the average number of transmissions, which depends on *r* the modified distance between the MTD and the UE (the distance taking into account path loss and shadowing effect).

Our analytical models were validated by simulation. In our simulations, all the devices are uniformly and independently distributed (homogeneous PPP) in a square area of 3 km x 3 km. Simulation parameters are summarized in Table 4-9. For the MTD power consumption in Tx, Rx and idle states, the same assumptions as in section 2.1 were taken. Same modulation and coding schemes (MCS) as in [LZG04] were considered, as well as the same signal to interference ratio (SIR) threshold.

Table 4-9 Simulation parameters

Parameter	Assumption
MTD power consumption in Tx state ($P_{m,T}$)	925 mW
MTD power consumption in Rx state ($P_{m,R}$)	200 mW
MTD power consumption in the idle state ($P_{m,I}$)	10 mW

MTD bandwidth	1.4 mHz
Packet size (L)	1080 bit
Path loss exponent	4
Density of active MTDs per m^2	16x10^{-6}
Standard deviation for log-normal shadowing	8 dB

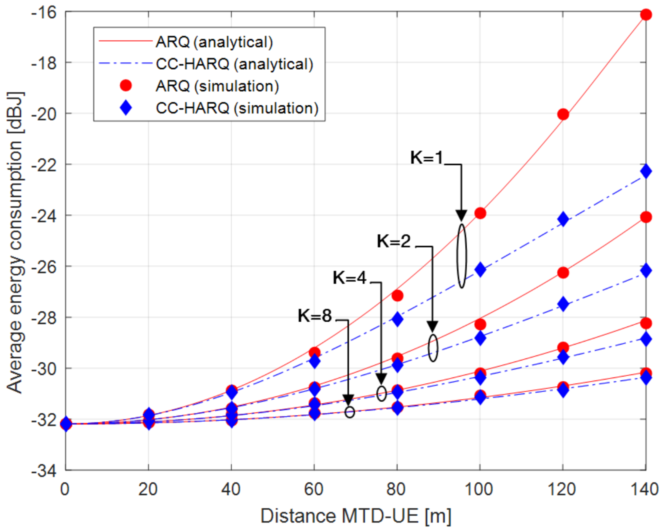


Figure 4-38 Average energy consumption of ARQ and CC-HARQ as a function of the modified distance MTD-UE, considering a QPSK modulation with a coding rate ¾, a 1.5 bit/sym. rate and a 4 dB SIR threshold, maximum number of transmissions N=128, and number of sub-channels K = {1; 2; 4; 8}

One of the main characteristics of the mMTC applications is the high connection density (devices/km^2), while one of the advantages of D2D relaying is the reuse of resources. When increasing the number of sub-channels shared by the MTDs, we can reduce the density of devices sharing a sub-channel. Figure 4-38 shows the average MTD energy consumption considering 1, 2, 4 and 8 sub-channels. From this figure, we observe that CC-HARQ outperforms ARQ in terms of energy consumption especially at a high density of MTDs sharing a sub-channel. For one sub-channel the average energy consumption with ARQ scheme can be up to four times higher than the average energy consumption with CC-HARQ scheme. Extended results are reported in [AVS+18]. Numerical results show that CC-HARQ outperforms ARQ in terms of energy consumption, especially for low relay density when the distance between an MTD and its relay increases or when the density of MTDs sharing the same sub-channel increases.

4.3.2.3 A D2D discovery protocol inspired by 802.11 RTS-CTS protocol for optimizing energy consumption

D2D discovery phase is a key enabler for D2D communications. In this phase an MTD looks for potential relays (UEs), and then, according to established criteria, it selects only a single relay among them. In mMTC applications, the discovery and the relay selection process must be low power consumption and low complexity taking into account the specific constraints of MTDs. For this purpose, we propose to join the discovery and the relay selection in a single process inspired by the 802.11 RTS/CTS (Request-to-Send/Clear-to-Send) protocol due to its simplicity and limited signal overhead.

Network model

We consider a single cell in which UEs and MTDs are randomly distributed. The locations of UEs form a homogeneous Poisson point process (PPP) Φ_u with density λ_u in $\mathbb{R}^2$ (see Figure 4-39). All the devices (MTDs and UEs) are in down-link coverage. MTDs with an unfavorable link budget have up-link coverage at the cost of more energy consumption. In order to reduce the MTD energy consumption, the MTD may use a nearby UE as relay. This approach consists of three phases: (i) the MTD selects a nearby UE as relay (D2D discovery), (ii) the MTD transmits its data to the selected relay (D2D communication), and (iii) the selected relay transmits the MTD data to the BS (cellular communication). However, we do not consider the third phase in our analysis since the MTD does not participate in this phase.

We consider that the BS allocates dedicated resources for D2D links (i.e., there is no interference between D2D links and cellular links). In order to keep the analysis tractable, we make the following assumptions:

1) Distance-based path loss (i.e. no fading, no shadowing).

2) Devices (MTDs and UEs) transmit at a fixed transmission power and have a bandwidth of Bω Hz.

3) Reciprocal channel (i.e., the channel from point A to B is the same as the channel from B to A).

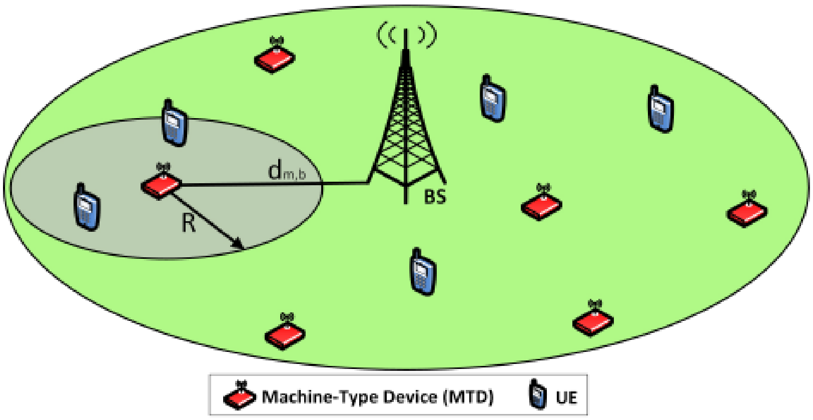


Figure 4-39 Network model

Protocol description:

1. The MTD measures the path loss value $L_{m,b}$ between itself and the BS. Based on this value, the MTD defines the contention window W and the path loss threshold between itself and a nearby UE (L_{th}). The MTD broadcasts an Request-for-Relay (RR) packet that carries the values of W and L_{th}. Then, the MTD waits for a response from a UE (contention process). The maximum waiting time is W time slots.
2. A UE that receives an RR packet measures the path loss between the MTD and itself ($L_{m,u}$) by comparing the received and transmission power. Only the UEs having $L_{m,u} < L_{th}$ participate in the contention process (these UEs are called candidate relays). Each candidate relay responds to the MTD with a Relay-Candidate (RC) packet. To send its RC packet, the candidate relay chooses randomly and uniformly one time slot $s \in [1, W]$,that is a model based on a uniform random choice of time slot. The first candidate relay that transmits a RC packet without collision wins the contention process.
3. The MTD broadcasts the feedback packet as soon as it successfully receives a RC packet. The feedback packet serves to confirm the winning UE and to avoid the hidden node problem. The hidden node problem occurs when a candidate relay is not able to hear the RC packet sent by another one. The candidate relay that wins the contention is called the selected relay. The other candidate relays stop contending when they hear the feedback packet.

In the case of no UE answers, it means that the path loss between the MTD and the BS is less than the path loss between the MTD and a UE. Therefore the MTD must continue with the direct transmission of its data towards the BS.

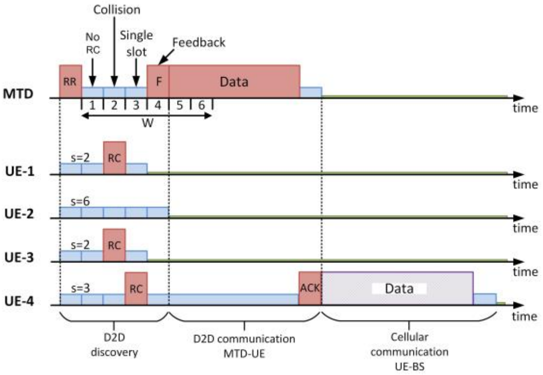


Figure 4-40 Packet exchange sequence in the D2D relaying mechanism when the MTD looks for and find a relay

Figure 4-40 shows the packet exchange sequence in a network where $W = 6$. We then used stochastic geometry to analytically derive the relay discovery probability, the number of slots used in the contention process and the total MTD energy consumption during a direct cellular communication mode and a D2D mode [AVS+19]. We have shown that the discovery probability can be maximized and the length of the contention window can be minimized so that the discovery area and the MTD energy consumption are optimized. Details of the analytical derivations are given in [AVS+19].

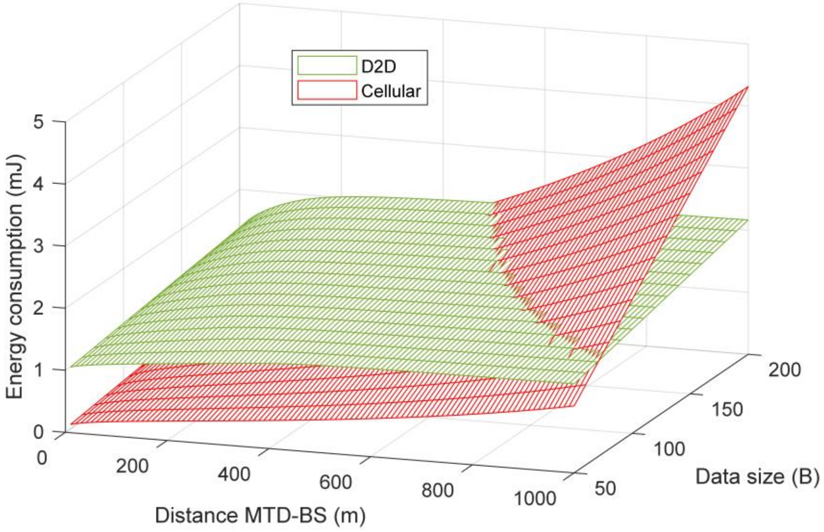


Figure 4-41 Comparison between the average energy consumption in cellular and in D2D mode as a function of the distance MTD-UE ($d_{m,b}$) and the data size, considering $\lambda_u = 100\ x\ 10^{-6}\ UEs/m^2$.

In Figure 4-41 we compare the average energy consumption in cellular mode and the minimum average energy consumption in D2D mode (i.e. considering the optimal discovery area minimizing the total energy consumption in the D2D mode). This figure shows that the D2D relay mechanism allows reducing the energy consumption significantly when the data packet size is relatively large compared to the discovery packet size (RR,RC, and feedback), and when the MTD is far from the BS.

With the parameters used in our simulations we have shown a clear improvement of the MTD energy consumption using D2D relaying (up to 50% gain) when the data packet size is larger than 150 bytes and when the MTD is far from the BS (unfavourable link budget).

4.4 Summary

Table 4-10 Summary of key recommendations and benefits in terms of spectrum, connectivity and mobility optimizations

Feature	Recommendation	E2E / KQI benefits
Spectrum optimization techniques		
Radio resource allocation strategies for services mapping	In a heterogeneous scenario, use a multi-cell aggregation scheduler (MAS) which, relying on the UE's CSI, determines what cell allows spectral efficiency and delay to be optimized when a UE is attached to it.	Expected improvement of E2E throughput and delay.
Dynamic spectrum aggregation for 5G new radio	Use dynamic bandwidth parts with/without the same central frequency for bandwidth adaptation and load balancing purposes, respectively.	Expected improvement of E2E throughput and delay (integrity).
LAA signaling assessment of eMBB services in unlicensed band (5Ghz)	Use DRS 160 ms periodicity Activate DRS compensation method (CDRS) when disabling DRS signals Automatic per service and load conditions priorities selection	In indoor ultra-dense scenarios up to 40% improvements in File Transfer Delay and 5% File Transfer Throughput while improving fairness in 15% towards WiFi.
Unlicensed standalone operation with MF	Achieving low latency communication in the 5 GHz unlicensed band is challenging due to the LBT procedures. For the uplink, GUL is recommended for low latency traffic over SUL.	Results show that use cases requiring one-way radio latency in the order of 30-40 ms with 99.9% reliability can be supported in the 5GHz band with MF. Using grant-free uplink transmissions offers 25% latency improvement at low to medium loads. Enabling K-repetition for ACK/NACK feedback gives ~20% latency improvement at low to medium loads. Omitting Cat 1 LBT during DL-2-UL transition results in up to 55% latency improvement when the offered load is high.
NR-U standalone	NR-U offers significant latency/reliability benefits as compared to MF due to shorter TTIs, more flexible frame structure, reduced gNB and UE processing times. But, LBT procedures still limit the latency budget.	Latencies of 8-17 ms at 99.99% reliability can be supported in the 5GHz band with NR-U.
Advanced mobility optimizations		

Social events information gathering, association and application to cellular networks	Social data is required to properly forecast and avoid service degradations	The increases in demand related to events highly impact the service provision. Social-aware optimization mechanisms allow for detecting social events as the cause of past degradations, forecasting of future increases of demand and load-balancing mechanisms allowing a 44% reduction in the peak increment of users served by the site closest to the venue.
QoE proactive management	Use a predictive framework for network performance forecast, so that occasional performance issues leading to UE's QoE degradation can be avoided by means of a proactive network configuration.	All the E2E KQIs can benefit from this research line. As a first step, however, integrity-related KQIs will be addressed, such as the E2E throughput.
Algorithm on mobility and access management	Use channel quality (CQI) measurements, location information and availability of connection to a cell to drive users' mobility	Expected improvement of E2E throughput and delay (integrity).
Utilization of Prediction in Small-Cell Mobility	PNC controlled strategy to forward PDCP PDUs in a MC connectivity scenario with a split bearer architecture. The PNC strategy will be based on periodic CSI reports captured in the PDCP layer in the MgNB which is acting as the base station for the primary cell (PCell).	Gains of 10% of throughput during the execution of a SCell change process in an URLLC client performing a detected trajectory over a set of SmallCells.
RRC State Selection for URLLC V2X	Exploitation of RRC Idle and RRC Inactive states for V2X applications	The increased number of devices impacts the V2X applications with strict delay requirements. The use of platoons and different RRC states can be used to allow for more non-V2X devices to be served, while respecting the stringent delay requirements of V2X.
MEC-assisted C-V2X Communications	Exploitation of MEC deployments, where edge hosts are co-located with radio connectivity nodes.	E2E latency reduction, as compared to "legacy" network architecture; such a reduction can be proven life-saving for critical scenarios such as the one of VRU.
Basic 5G NR mobility solutions	Use synchronous RA-less handovers. Use network-assisted UE autonomous secondary cell management. Use multi-node connectivity when feasible for achieving zero	Offers reduced handover interruption times, enhanced mobility robustness with low HOF and RLF rates, and reduced signaling (RRC, RA, and Xn) overhead.

	handover interruption times and enhanced robustness. Conditional handovers for selected services.	Maps to improvements in KQI service retainability and service integrity.
Connectivity optimizations for device-to-device communications and relaying paths		
Stochastic resource optimization for heterogeneous architecture	The optimal distributed scheduling that achieves a tradeoff between total throughput and energy consumption can be obtained by an appropriate modification of the CSMA/CA.	3-5 times less power consumption in average, improvement of E2E throughput.
Power consumption reduction for mMTC	Binary power control (i.e. on/off with max power) achieves a tradeoff between minimizing power consumption and average delay for low traffic arrival and Gilbert-Eliot channel.	Improvement of latency by 10% and service retainability (battery life).
D2D relaying for eMBB	In the scenario of limited CSI feedback, relay selection made at the user side achieves a better queuing stability region as compared to the case where the relay selection is made by the BS.	Improved queuing stability, which results in improved E2E throughput, by 25%.
D2D relay mechanism for mMTC services	Use the autonomous resource allocation mode and consider using CC-HARQ as retransmission scheme to take care of interference and increase the transmission success probability Use our RR/RC protocol as low complexity D2D discovery protocol	Optimization of resource allocation Optimization of the energy consumption (up to 50% gain for 200 bytes packet size and devices located beyond 500 m from the BS) Low complexity for the MTC device Improved KQIs network accessibility and service retainability (battery life)

5 Impact on 3GPP NR standardization

ONE5G WP3's impact on 3GPP NR standardization is summarized in the following. Table 5-1 summarizes impact on the NR Release-15 specifications. In this context, it should be noted that the 3GPP entered the work item phase for completion of Release-15 at the start of ONE5G. Secondly,

Table 5-2 summarizes ONE5G WP3's impact on various 3GPP NR Release-16 Study and Work Items. NR Release-16 is set to be finalized by end of 2019. Finally, several of the developed innovations in this project are candidates to be included in future 3GPP NR releases such as 17 and 18, those are summarized in Table 5-3.

Table 5-1 Summary of impact on 3GPP NR Release-15 specifications

Feature	Description	3GPP reference
Preemptive scheduling	Mux of eMBB and URLLC with different TTI sizes, interrupted transmission indication and CBG-based HARQ retransmissions.	3GPP TS 38.300 Section 10.2, 3GPP TS 38.214 Sections 9.1 and 11.2.
RRC state machine	Definition of the new RRC INACTIVE state and related state transition rules.	3GPP TS 38.300 Section 7, 3GPP TS 38.304, and 3GPP TS 38.331
FR1 & FR2 Carrier Aggregation	Radio resource allocation for services mapping, needs the possibility of aggregating NR FR1 (N77) & NR FR2 (N258) simultaneously	3GPP TS 38.101 Section 5.2 and TS 138.133 Section 8.1.7
UE-specific BWP inactivity timer configuration	UE-specific BWP inactivity timer can be configured for each serving cell so as to optimize the UE operation time duration in the configured non-default BWPs to enable power saving.	3GPP TS 38.321 Section 5.15 and TS 38.331 Section 6.3.2

Table 5-2 Summary of impact on 3GPP NR Release-16 Study and Work Items

Feature	Description	3GPP reference
UE power consumption model and WUS-triggered DRX	State-of-the-art UE power consumption model, capturing effects of different RRC states and DRX, and WUS-triggered DRX study.	3GPP SI on UE power consumption (RP-181463) and 3GPP TR 38.840.
Enhanced DC/CA with PDCP duplication	Resource efficient DC/CA with PDCP data duplication for URLLC services and its extension to up to 4 duplicates.	3GPP SI on IIoT (RP-181479) and 3GPP TR 38.825, and 3GPP WI on IIoT (RP-190728).
Enhanced mux of eMBB and URLLC	Inter-UE mux of eMBB and URLLC and URLLC system-level performance assessment	3GPP SI eURLLC (RP-181477) and 3GPP TR 38.824.
DPS and Centralized URLLC scheduling	Evaluation of multi-TRP DPS and centralized multi-cell scheduling of URLLC.	MIMO (multi-TRP aspects) (RP-182075)
FR1 (NR) & FR2 (NR) Carrier Aggregation	Radio resource allocation for services mapping, needs the possibility of aggregating FR1 (N77) & FR2 (N258) simultaneously	3GPP WI on DC and CA enhancements (RP 181469)
In-band signaling enhanced DRS mechanism for the operation in unlicensed bands	Proposed DRS signaling modifications in dense coexistence scenario for improved eMBB service performance measured in KQI	3GPP New WID on NR-based Access to Unlicensed Spectrum (RP-182806)

QoE-KQI-KPI mapping	This functionality gives response to the difficulty of network operators to gather high-layer end-to-end performance metrics (KQIs and QoE) by estimating them from lower layers metrics using machine learning techniques. These high-layer performance metrics will afterwards be used by other network management mechanisms.	3GPP Study of enablers for Network Automation for 5G (TR 23.791).

For future 3GPP NR Releases (i.e. Rel-17 and Rel-18), the following table summarizes the developed enhancements that ONE5G consider as promising for upcoming standardization activities.

Table 5-3 Candidate features for future 3GPP NR Releases such as 17 and 18.

Feature	Description	3GPP impact
MU-MIMO null-space preemptive scheduling	Promising method that exploits the spatial dimension to more efficiently multiplexing eMBB and URLLC users for cases with at least 8 gNB antennas.	Requires additional NR standardization of gNB-2-UE signaling to facilitate good isolation between co-scheduled eMBB and URLLC users (i.e. new DCI formats)
Enhanced C-RAN multi-cell scheduling	Centralized multi-cell scheduling offers significant benefits. Developed methods can to a large extended be implemented for NR Release 15 and 16 specs, but additional 3GPP specs would still be useful.	Enhanced signaling options for the F1 (higher layer split option) and F2 (lower layer split option) interfaces, should 3GPP chose to standardize F2.
a D2D relaying scheme for mMTC	A D2D relaying scheme adapted to the specific mMTC constraints in terms of energy consumption and based on a discovery protocol inspired by the 802.11 RTS/CTS protocol. Promising performances in terms of optimization of the MTC device energy consumption have been shown.	A new discovery procedure to be included in D2D work item for NR mMTC rel. 17/18
Component carrier management in multi-connectivity environment	This functional block aims at dynamically assign PSCells and SCells to UEs, by hosting these in the CCs that best fit according to network operators' policies. This will bring benefits both in terms of enhanced throughput and in reliability, depending on whether the data flow among the selected CCs is split or duplicated, respectively.	- New mechanism for cell addition/removal/change (beyond current mobility triggering events). - Development of mechanisms to gather/process network performance information beyond traditional radio KPIs (e.g., context, KQIs, etc.). Initially addressed in Rel-16 TR 23.791 with NWDAF.
Uneven traffic split among component carriers in multi-connectivity environment	This functionality aims at complementing the dynamic assignment of component carriers and provides a way to fine-tune the usage of radio resources among currently assigned component carriers, according to the network state by appropriately assigning the amount of traffic to be held by each of them.	- Currently applicable to 5G dual-connectivity scenarios. - However, for its benefits to be maximized, it should be operated jointly with the dynamic component carrier management in a 5G multi-connectivity scenario.
QoE steering	Following a mobility-based approach, this mechanism uses UE and network performance information to estimate each UE's perceived QoE, which is	- Inclusion/development of methods for per-service QoE estimation using lower-layers performance information.

	afterwards used to hand UEs over cells in order to steer/balance this QoE.	- Inclusion of per-service handover margins.
Proactive context-aware network management	This functionality aims at analysing and forecasting network performance, identifying the cause of past degradations and proactively identifying future ones before they actually have taken place, particularly for those generated by causes outside of the network elements themselves (e.g. event-caused crowds). In this way, corrective actions can be defined and applied in advance (e.g. preventive allocation of resources). This will allow preventing UEs from experiencing such degradations and an optimized allocation of resources.	- Development of mechanisms to gather/process context information from a variety of sources (e.g., social networks). Initially addressed in Rel-16 TR 23.791 with NWDAF.
Dynamic resource allocation for service mapping	Objective is to specify radio resource management algorithms that take into account the service policies by allocating dynamically the service to the macro gNodeB and/or µυgNodeB	New mechanism to select/aggregate/connect the cells to optimize users' resource allocation in dense network
Small Data Transmission in RRC INACTIVE state	Design of small data transmission (SDT) during RRC INACTIVE state, without state transition to RRC CONNECTED.	Requires additional NR standardization of e.g. RRC and MAC related to the UE procedures in RRC INACTIVE.
RRC state handling for URRL V2X	Objective is to assess the benefits of using the idle and inactive states for V2X applications, taking into account the stringent delay requirements.	Framework for state selection between idle and inactive in V2X applications.
Dynamic resource allocation for URLLC services	Design of a resource allocation policy with absence of CSI knowledge at the transmitter	Adapted for cases where the BS cannot know the CSI of the URLLC users due to the short latency. In future releases, with the increase of number of URLLC users with more stringent latency requirements, the proposed method will be interesting.
Stochastic resource optimization for heterogeneous architecture	We developed fully distributed scheduling for D2D that achieves a trade-off between throughput and power consumption.	This requires new signaling and frame structure since the proposed policy requires the implementation of a contention procedure.
Power consumption reduction for mMTC	Development of a promising framework that has led to a simple transmission strategy for mMTC (on-off) that reduces their power consumption	The method requires a new signaling in both uplink and downlink
D2D relaying for eMBB	Development of a new distributed relay selection policy in the context of limited feedback.	The framework requires new signaling to send the CSIs and the relay selection in D2D.
Configured grant assignment for misaligned periodic traffic	Proposed scheme to continually adjust BS' estimate of traffic periodicity and time of arrival of next packet.	The scheme requires that the BS keeps and updates simple state variables when observing new arrivals.

6 Conclusion

In this WP3 deliverable we have presented our final recommendations regarding the technologies and innovations studied in WP3 throughout the ONE5G project with the aim of optimizing the E2E performance of the 3GPP 5G NR. The developed enhancements are largely generic in the sense that they are applicable to both the considered Megacity and Underserved scenarios. It is noted that the E2E performance benefits of each innovation are defined based on the KQI framework that allows to highlight the E2E aspects. The proposed solutions were validated by a mixture of semi-analytical and heuristic methods, including examples of proof of optimality for selected cases, wherever the derivation of such proofs was feasible. Tools from classical optimization theory and machine learning discipline have also been widely utilized. Throughout the project duration, the NR system design principles and performance assessment assumptions have been adopted to closely follow the 3GPP guidelines. In short, various control plane and terminal energy consumption optimizations were presented in Chapter 2. In Chapter 3 we presented various multi-service scheduling solutions and innovations for different service mixtures (e.g. URLLC and eMBB services) and network architectures (D-RAN, C-RAN). Chapter 4 covered E2E optimizations related to spectrum, mobility and connectivity. At the end of each of the Chapters 2-4, we reported a table summary of the developed E2E optimization enablers, highlighting the recommendations on how and in which scenario(s) to benefit from a proposed scheme and the anticipated E2E benefits (quantified numerically for many cases). Those performance benefits were mainly obtained from advanced system-level simulations and also some semi-analytical results based on stochastic geometry models. In Chapter 5, we presented a summary of how a selected set of the developed E2E performance features are linked to 3GPP standardization of NR. From Chapter 5 it is visible that ONE5G has had some early links to 3GPP NR Rel-15 (a.k.a. the first 5G standards release), and has conducted research resulting in impacting the ongoing NR Rel-16 standardization process, as well as has developed a promising set of forward-looking features that may be considered for future NR Rel-17 and Rel-18 standardization activities.

In short, the WP3 main research findings can be condensed to the following five take-aways:

1) Control plane and UE energy consumption are important for E2E performance optimizations. The new three-state RRC machinery for NR offers new opportunities for optimizing the tradeoffs between latency, signalling overhead, and UE energy. ONE5G has developed recommendations how to best utilize these new degrees of freedom, including the utilization of the enhanced DRX framework and so-called BWP mechanisms. As an example, up to 89% shorter CP latency at transition from RRC_Inactive to Connected (compared to Idle) is achieved, with ~40-70% longer battery life of RRC_Inactive compared to Connected for scenarios with infrequent traffic. Only ~10% latency increase of RRC_Inactive compared to Connected for infrequent traffic. Concepts and building blocks for device virtualization schemes have been presented as well in a setting with centralized network architecture. Details are covered in Chapter 2.
2) Efficient multi-user and multi-service resource allocation is also of high importance for E2E performance optimization, a.k.a. scheduling. A promising set of novel scheduling solutions has been developed, including support for cases with diverse sets of services, cases with distributed and centralized network architectures, single- and dual-node connectivity, etc. The reported findings in Chapter 3 (and in the related quoted publications from ONE5G) confirm the promising gains of the developed solutions. Gains range from approximately 20%-50% and up to more than 100% for advanced cases with either multi-node connectivity and/or C-RAN architectures.
3) Efficient spectrum utilization is obviously of great importance. In this domain, WP3 has developed enhanced KQI-based spectrum selection methods, including aggregation of different licensed/unlicensed spectrum chunks that include same or different RATs, and also stand-alone unlicensed spectrum operation based on Multefire and NR-U. Those spectrum utilization enhancements are found to offer significant benefits of

approximately 40% (but naturally varies depending on the exact case), as compared to known solutions. Achieving low latency with high reliability is found to be particularly challenging for unlicensed bands (5GHz) due to the significant amount of time spend on LBT procedures. As an example, latencies of 8-17 ms with 99.9% reliability are found to be realistic with NR-U. See more details in Chapter 4.

4) Service-dependent traffic steering mechanisms, driven by machine learning techniques, with the aim to achieve novel QoE-driven balancing solutions (as compared to traditional load balancing) are found to be very promising, showing gains on the order of 30%-60%. Furthermore, social events information gathering, association and application to learning based prediction in cellular network performance data have been developed and tested on data from real networks. Results show that using such techniques is promising for achieving E2E performance benefits. See more details in Chapter 4.
5) Several novelties were developed for D2D connectivity cases: an innovative resource optimization for heterogeneous architecture helps achieving good performance for delay-restricted applications, while minimizing power consumption, a D2D relay method for eMBB services accomplishes better queuing stability, and a D2D relay mechanism to jointly improve mMTC accessibility (measured in KQI terms) and energy consumption has been introduced. As an example, power consumption reductions of a factor 3-5 are achieved, as well as 25%-50% throughput improvements. See more details in Chapter 4.

Finally, some contributions described in this deliverable have been implemented and validated using the common system level simulator platform for the ONE5G project developed in WP2 and others are demonstrated by Proof-of-Concept (PoC) as part of the WP5 activities. Findings from WP3 have been widely published at internationally recognized conferences and prestigious journals (see the many references throughout D3.2).

References

[3GPP] 3GPP website, http://www.3gpp.org/

[3GPP22.261] 3GPP, “TS 22.261 service requirements for next generation new services and markets,” 3rd Generation Partnership Project (3GPP), v16.3.0, 2018.

[3GPP24.301] 3GPP, “TS 24.301 Universal Mobile Telecommunications System (UMTS); LTE; 5G; Non-Access-Stratum (NAS) protocol for Evolved Packet System (EPS); Stage 3” 3rd Generation Partnership Project (3GPP), v15.3.0, 2018.

[3GPP28.801] "Study on management and orchestration of network slicing for next generation network (Release 15)", January 2018

[3GPP36.205] 3GPP, “TS 36.205 Evolved Universal Terrestrial Radio Access (E-UTRA); Physical layer procedures”), v14.8.0, 2018.

[3GPP36.305] 3GPP, “TS 36.305 Evolved Universal Terrestrial Radio Access Network (E-UTRAN); Stage 2 functional specification of User Equipment (UE) positioning in E-UTRAN,” 3rd Generation Partnership Project (3GPP), v15.1.0, 2018.

[3GPP36.839] 3GPP Technical Report 36.839, “Mobility Enhancements in Heterogeneous Networks”, available at www.3gpp.org.

[3GPP36.843] 3GPP, “Study on LTE Device to Device Proximity Services; Radio Aspects,” 3rd Generation Partnership Project (3GPP), TR 36.843 V12.0.1, March 2014. available at www.3gpp.org.

[3GPP36.888] 3GPP, “Study on provision of low-cost Machine-Type Communications (MTC) User Equipments (UEs) based on LTE,” 3rd Generation Partnership Project (3GPP), TR 36.888 V12.0.0, June 2003. available at www.3gpp.org

[3GPP36.889] 3GPP TR 36.889 v13.0.0, “Feasibility Study on Licensed-Assisted Access to Unlicensed Spectrum”, July 2015; www.3gpp.org.

[3GPP36.912] 3GPP TR 36.912 v15.0.0, “Feasibility study for Further Advancements for E-UTRA (LTE-Advanced)”, June 2018; www.3gpp.org.http://www.3gpp.org/

[3GPP37.240] 3GPP, “TS 37.240 Evolved Universal Terrestrial Radio Access (E-UTRA) and NR; Multi-connectivity; Stage 2 (Release 15),” 3rd Generation Partnership Project (3GPP), v15.0.3, 2018.

[3GPP37.340] 3GPP TS 37.340 Evolved Universal Terrestrial Radio Access (E-UTRA) and NR; Multi-connectivity; Stage 2 (Release 15)

[3GPP38.101] 3GPP TS 38.101-1, User Equipment (UE) radio transmission and reception; Part1: Range 1 Standalone (Release 15); V15.2.0 (2018-06)

[3GPP38.133] 3GPP TS 38.133, Requirements for support of radio resource management (3GPP TS 38.133 version 15.2.0 Release 15)

[3GPP38.211] 3GPP Technical Specification 38.211, “Technical Specification Group Radio Access Network; NR; Physical channels and modulation”, available at www.3gpp.org.

[3GPP38.300] 3GPP TS 38.300 NR; NR and NG-RAN Overall Description; Stage 2 (Release 15)

[3GPP38.425] TS 38.425, NG-RAN; NR user plane protocol, v.15.3,0, September 2018

[3GPP38.801] 3GPP Technical Report 38.801, “Technical Specification Group Radio Access Network; Study on new radio access technology: Radio access architecture and interfaces”, available at www.3gpp.org.

[3GPP38.840] 3GPP TR 38.840, Study on UE power saving in NR, March 2019

[3GPP38.889] 3GPP TR 38.889 v16.0.0, “Study on NR-based access to unlicensed spectrum”, December 2018; www.3gpp.org.

[3GPP38.803]	3GPP TR, Study on new radio access technology: Radio Frequency (RF) and co-existence aspects, v.14.2.0, September 2017
[3GPP38.913]	3GPP TR 38.913 V14.1.0 (2016-12) - 3rd Generation Partnership Project; Technical Specification Group Radio Access Network; Study on Scenarios and Requirements for Next Generation Access Technologies; (Release 14)
[3GPPRAN]	3GPP RAN1#94b Chairman's note
[3GPPR1-1811478]	Evaluation methodology for UE power consumption estimation, Nokia, October 2018.
[3GPPR2-1817582]	3GPP Technical Document R2-1817582, "Resource Efficient PDCP Duplication," Nokia, November 2018.
[3GPPR3-186693]	3GPP Technical Document R3-186693, "Selective duplication upon transmission failure," Nokia, November 2018.
[3GPPR3-185547]	3GPP Technical Document R3-185547, "Resource Efficient PDCP Duplication," Nokia, October 2018.
[3GPPRP-181463]	Study on UE Power savings in NR, June 2018.
[3GPPRP-190728]	3GPP RP-190728, Support of NR Industrial Internet of Things (IoT), Nokia, March 2019
[3GPPRP-182090]	3GPP RP-182090, Study on NR Industrial Internet of Things (IoT), Nokia, September 2018.
[3GPP_NRU]	3GPP RP-181339, "Study on NR-based Access to Unlicensed Spectrum", June 2018; available at www.3gpp.org
[5GAA17]	"Toward fully connected vehicles: Edge computing for advanced automotive communications," 5G Alliance for Connected Industries and Automation, Tech. Rep., 2017.
[5GAA18]	"5G for connected industries and automation," 5G Alliance for Connected Industries and Automation, Tech. Rep., 2018.
[AKR+14]	Akramullah, Shahriar. "Digital Video Concepts, Methods, and Metrics", 2014
[ASP14]	A. Alvarado, G. Scutari and J. Pang, "A New Decomposition Method for Multiuser DC-Programming and Its Applications," in *IEEE Transactions on Signal Processing*, vol. 62, no. 11, pp. 2984-2998, June, 2014.
[AVS+18]	C. Vargas Anamuro, N. Varsier, J. Schwoerer and X. Lagrange, "Modeling of MTC Energy Consumption for D2D communications with Chase Combining HARQ Scheme," 5GNR workshop in Globecom, Abu Dhabi, Dec. 2018.
[AVS+19]	C. Vargas Anamuro, N. Varsier, J. Schwoerer and X. Lagrange, "Performance analysis of a discovery process for UE-relayed MTC communications," submitted to the 12th IFIP Wireless and Mobile Networking conference, Paris, 2019.
[BCG+17]	B. Chen, J. Chen, Y. Gao and J. Zhang, "Coexistence of LTE-LAA and Wi-Fi on 5 GHz With Corresponding Deployment Scenarios: A Survey," in IEEE Communications Surveys & Tutorials, vol. 19, no. 1, pp. 7-32, 2017.
[BKP+16]	G. Berardinelli, S. R. Khosravirad, K. I. Pedersen, F. Frederiksen and P. Mogensen, "Enabling Early HARQ Feedback in 5G Networks," inProc. IEEE VTC Spring, Nanjing, China, pp. 1-5, 2016.
[BAD19]	N. Ben Khalifa, M. Assaad, M. Debbah, "Risk-sensitive Reinforcement Learning for URLLC Traffic in Wireless Networks," in proc. of IEEE WCNC, Marrakesh, 2019.

[BMS+12a] S. Barbera, P. Michaelsen, M. Saily, K.I. Pedersen, “Improved Mobility Performance in LTE Co-Channel HetNets Through Speed Differentiated Enhancements”, in IEEE Proc. Globecom, December 2012.

[BMS+12b] S. Barbera, P.H. Michaelsen, M. Sailly, K.I. Pedersen, “Mobility Performance of LTE Co-Channel Deployment of Macro and Pico Cells”, in IEEE Proc. WCNC, April 2012.

[BPR+15] S. Barbera, K.I. Pedersen, C. Rosa, P.H. Michaelsen, F. Frederiksen, E. Shah, A. Baumgartner, ”Synchronized RACH-less Handover Solution for LTE Heterogeneous Networks, in IEEE Proc. International Symposium on Wireless Communication Systems (ISWCS), August 2015.

[BRE01] L. Breiman, “Random Forests”, Machine Learning, 45(1), 5-32, 2001.

[CDH17] Alexandre Cebeillac, Éric Daudé et Thomas Huraux, « Where ? When ? And how often ? What can we learn about daily urban mobilities from Twitter data and Google POIs in Bangkok (Thailand) and which perspectives for dengue studies ? », Netcom, 31-3/4 | 2017, 283-308.

[CHASE85] D. Chase, “Code combining: A maximum-likelihood decoding approach for combining an arbitrary number of noisy packets,” IEEE Transactions on Communications, vol. 33, no. 5, 1985.

[CK10] Sven F. Crone, Nikolaos Kourentzes, “Feature selection for time series prediction – A combined filter and wrapper approach for neural networks”, in Neurocomputing, vol. 73(10-12), pp. 1923-1936, 2010.

[CMG+14] Cho, Kyunghyun; van Merrienboer, Bart; Gulcehre, Caglar; Bahdanau, Dzmitry; Bougares, Fethi; Schwenk, Holger; Bengio, Yoshua (2014). "Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation". arXiv:1406.1078

[DAD+18] M. Deghel, M. Assaad, M. Debbah, A. Ephremides, “Traffic-Aware Scheduling and Feedback Reporting in Multichannel Wireless Networks with Relaying”, journal paper, submitted, 2018.

[EE13] A. Elnasher, M.A. El-Saidny, “Looking at LTE in Practice: A performance analysis of the LTE system based on field results”, in IEEE Vehicular technology Magazine, September 2013.

[EFK18] M. Emara, M.C. Filippou and K. Ingolf, “Availability and Reliability of Wireless Links in 5G Systems: A Space-Time Approach”, in 2018 IEEE Global Communications URLLC Workshop

[EFS18] M. Emara, M. C. Filippou and D. Sabella, "MEC-Assisted End-to-End Latency Evaluations for C-V2X Communications," 2018 European Conference on Networks and Communications (EuCNC), Ljubljana, Slovenia, pp. 1-9, 2018.

[EP18a] A. A. Esswie, and K.I. Pedersen, “Multi-user preemptive scheduling for critical low latency communications in 5G networks,” in Proc. IEEE ISCC, Natal, pp. 1-6, 2018.

[EP18b] A. A. Esswie, and K.I. Pedersen, “Null space based preemptive scheduling for joint URLLC and eMBB traffic in 5G networks,” in Proc. IEEE Globecom, Abu Dhabi, Dec. 2018.

[EP18c] A. A. Esswie and K. I. Pedersen, "Opportunistic spatial preemptive scheduling for URLLC and eMBB coexistence in multi-user 5G networks," IEEE Access., vol. 6, pp. 38451-38463, July 2018.

[EP18d] A. A. Esswie, and K.I. Pedersen, “Capacity Optimization of Spatial Preemptive Scheduling for Joint URLLC-eMBB Traffic in 5G New Radio,” in Proc. IEEE Globecom, Abu Dhabi, Dec. 2018.

[ETSI5GHz] “5 GHz RLAN; Harmonised Standard covering the essential requirements of article 3.2 of Directive 2014/53/EU”, ETSI EN 301 893 V2.1.1 (2017-05), https://www.etsi.org/deliver/etsi_en/301800_301899/301893/02.01.01_60/en_301893v020101p.pdf

[ETSISTQ] "Speech and multimedia Transmission Quality (STQ); QoS aspects for popular services in mobile networks; Part 2: Definition of Quality of Service parameters and their computation", ETSI TS 102 250-2 V2.6.1 (2017-10), http://www.etsi.org/deliver/etsi_ts/102200_102299/10225002/02.06.01_60/ts_10225002v020601p.pdf

[EU2017] "COMMISSION IMPLEMENTING DECISION (EU) 2017/1483 of 8 August 2017", Official Journal of the European Union, https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32017D1483&from=EN

[Fan5GD4.2] FANTASTIC5G, Deliverable 4.2, "Final Results for the flexible 5G Air Interface multi-node/multi-antenna solution", April 2017.

[FPS+18] S. Fortes, D. Palacios, I. Serrano and R. Barco, "Applying Social Event Data for the Management of Cellular Networks," in IEEE Communications Magazine, vol. 56, no. 11, pp. 36-43, November 2018.

[FSB+17] S. Fortes, I. Serrano, R.Barco, "Cellular Network Management Based on Automatic Social-Data Acquisition", filled on May 2017, PCT/EP2017/060312.

[FAB+15] S. Fortes, A. Aguilar-García, R. Barco, F. B. Barba, J. A. Fernández-luque and A. Fernández-Durán, "Management architecture for location-aware self-organizing LTE/LTE-a small cell networks," in IEEE Communications Magazine, vol. 53, no. 1, pp. 294-302, January 2015.

[GB09] Gueguen, C., & Baey, S. (2009). A fair opportunistic access scheme for multiuser OFDM wireless networks. *EURASIP Journal on Wireless Communications and Networking*, *2009*, 14.

[GCS+16] L.C. Gimenez, M. Carmela; M. Stefan, K.I. Pedersen, A. Cattoni, "Mobility Performance in Slow- and High-Speed LTE Real Scenarios", in IEEE Proc. VTC-2016-Spring, May 2016.

[GMP16a] L.C. Giménez, P. H. Michaelsen, K. I. Pedersen, "UE Autonomous Cell Management in a High-Speed Scenario with Dual Connectivity", in 27th Annual IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC),September 2016.

[GMP16b] L. C. Gimenez, P. H. Michaelsen, K. I. Pedersen, "Analysis of data interruption in an LTE highway scenario with dual connectivity", in IEEE Proc. VTC-2016-Spring, May 2016.

[GTI17] GTI, "GTI Sub-6GHz 5G Device White Paper," 2017.

[HQG+12] J. Huang et al., "A close examination of performance and power characteristics of 4G LTE networks," in Proc. of MObiSys'12, UK, Jun. 2012.

[IEEE11ac] IEEE website, https://standards.ieee.org/standard/802_11ac-2013.html

[ITU-T] ITU-T TR GSTR-TN5G "Transport network support of IMT-2020/5G", February 2018.

[JW10] L. Jiang and J. Walrand, "A Distributed CSMA Algorithm for Throughput and Utility Maximization in Wireless Networks," IEEE/ACM Transactions on Networking, vol. 18, no. 3, pp. 960–972, June 2010.

[KGP17] T.E. Kolding, L. C. Gimenez, K.I. Pedersen, "Optimizing Synchronous Handover in Cloud RAN", in IEEE Proc. VTC-fall, September 2017.

[KLA18] S. Kriouile, M. Larranaga, M. Assaad, "Asymptotically Optimal Delay-aware Scheduling in Wireless Networks", submitted to IEEE Transactions on Information Theory, 2018.

[KLR19] A. Khlass, D. Laselva, and R. Järvelä, " Performance Enhancement with Efficient Radio Resource Control for 5G networks," submitted to IEEE PIMRC, Istanbul, Turkey, 2019.

[KMM17] U. Karneyenka, K. Mohta, & M. Moh, “Location and Mobility Aware Resource Management for 5G Cloud Radio Access Networks,” 2017 International Conference on High Performance Computing & Simulation (HPCS), 2017.

[KPM19] A Karimi, K. Pedersen, P. Mogensen, “5G NR URLLC Performance Analysis of DPS Frequency-Selective Multi-User Resource Allocation on URLLC in 5G NR”, submitted to IEEE Proc. ISWCS 2019.

[KPM+18] A. Karimi, K. I. Pedersen, N. H. Mahmood, J. Steiner, and P. Mogensen, “Centralized joint cell selection and scheduling for improved URLLC performance,” Accepted to be presented in 29th Annual IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC), September, 2018.

[KPM+19] A Karimi, K. Pedersen, N. Mahmood, G. Pocovi, P. Mogensen, “Efficient Low Complexity Packet Scheduling Algorithm for Mixed URLLC and eMBB Traffic in 5G”, in IEEE Proc. VTC-2019-spring, May 2019.

[LAD18] M. Larranaga, M. Assaad, K. De Turck, “Queue–aware Energy Efficient Control for Dense Wireless Networks“, in proc. of IEEE International Symposium on Information Theory (ISIT), 2018.

[LAU14] M. Lauridsen, “Studies on Mobile Terminal Energy Consumption for LTE and Future 5G”, Ph.D. Dissertation, Nov. 2014

[LBN13] M. Lauridsen, P. Mogensen and L. Noel, "Empirical LTE Smartphone Power Model with DRX Operation for System Level Simulations," 2013 IEEE 78th Vehicular Technology Conference (VTC Fall), Las Vegas, NV, pp. 1-6, 2013.

[LBS+14] M. Lauridsen, G. Berardinelli, T.B. Sorensen, P. Mogensen, “Ensuring Energy Efficient 5G User Equipment by Technology Evolution and Reuse”, in IEEE Proc. VTC-spring, May 2014.

[LKP+18] M. Lauridsen, T. Kolding, G. Pocovi and P. Mogensen, "Reducing Handover Outage for Autonomous Vehicles with LTE Hybrid Access," 2018 IEEE International Conference on Communications (ICC), Kansas City, MO, pp. 1-6, 2018.

[LMK+18] D. Laselva, M. Mattina, T. E. Kolding, J. Hui, L. Liu, and A. Weber, “Advancements of QoE Assessment and Optimization in Mobile Networks in the Machine Era,” in IEEE WCNC 2018, FlexNets workshop, April 2018.

[LZG04] Q. Liu, S. Zhou and G. B. Giannakis, “Cross-Layer combining of adaptive Modulation and coding with truncated ARQ over wireless links,” in IEEE Transactions on Wireless Communications, vol. 3, no.5, pp. 1746-1755, Sept. 2004.

[MAE18] A. Maatouk, M. Assaad, A. Ephremides, “Energy Efficient and Throughput Optimal CSMA”, submitted to IEEE Transactions on Networking, 2018.

[Mey07] S. P. Meyn, "Control Techniques for Complex Networks," Cambridge University Press, 2007

[MFA] MulteFire website, https://www.multefire.org/

[MLF+19] M. Lauridsen, D. Laselva, F. Frederiksen, J. Kaikkonen, "5G New Radio User Equipment Power Modeling and Potential Energy Savings," submitted to IEEE VTC Fall, Honolulu, Hawaii, USA, 2019.

[MLL+18] N. H. Mahmood, M. Lopez, D. Laselva, K. Pedersen and G. Berardinelli, "Reliability Oriented Dual Connectivity for URLLC services in 5G New Radio," 2018 15th International Symposium on Wireless Communication Systems (ISWCS), Lisbon, Portugal, pp. 1-6, 2018.

[MLP+18] N. H. Mahmood, D. Laselva, D. Palacios, M. Emara, M. C. Filippou, D. M. Kim, I. de-la-Bandera, "Multi-Channel Access Solutions for 5G New Radio," Submitted to IEEE Communications Magazine, October 2018.

[MRF+18a] R. Maldonado Cuevas, C. Rosa, F. Frederiksen and K.I. Pedersen, "On the Impact of Listen-Before-Talk on Ultra-Reliable Low-Latency Communications", to appear in the proceeding of IEEE Global Communications Conference 2018, December 2018.

[MRF+18b] R. Maldonado Cuevas, C. Rosa, F. Frederiksen and K.I. Pedersen, "Uplink ultra-reliable low latency communications assessment in unlicensed spectrum", to appear in the proceeding of IEEE Global Communications Conference 2018, December 2018.

[MRF+19] R. Maldonado Cuevas, C. Rosa, F. Frederiksen and K.I. Pedersen, "Analysis of Ultra-Reliable Low Latency Communication for Unlicensed Spectrum with Joint Uplink and Downlink Traffic", submission to IEEE Access (planned for April-2019).

[MVL+18] H. Martikainen, I Viering, A. Lobinger, T. Jokela "On the Basics of Conditional Handover for 5G Mobility", in IEEE Proc. International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC) 2018 , Bologna, Italy, September, 2018.

[NLS10] J. Navarro-Ortiz, J. M. Lopez-Soler and G. Stea, "Quality of experience based resource sharing in IEEE 802.11e HCCA," 2010 European Wireless Conference (EW), Lucca, 2010, pp. 454-461.

[NGMN16] Next Generation Mobile Networks Alliance, "Recommendations for NGMN KPIs and requirements for 5G," 2016.

[NGMN+18] NGMN Alliance "Overview on 5G RAN Functional Decomposition", February 2018.

[NLL+18] N. H. Mahmood, M. Lopez, D. Laselva, K. Pedersen, and G. Berardinelli, "Reliability Oriented Dual Connectivity for URLLC services," in Proceeding of ISWCS '18, Lisbon, Portugal, August 2018.

[NS3] Ns-3 website, https://www.nsnam.org/

[OTL18] P. Oliver-Balsalobre, M. Toril, S. Luna-Ramírez and R. G. Garaluz, "Self-Tuning of Service Priority Parameters for Optimizing Quality of Experience in LTE," IEEE Transactions on Vehicular Technology, vol. 67, no. 4, pp. 3534-3544, April 2018.

[ONE17-D21] ONE5G Deliverable D2.1, "Scenarios, KPIs, use cases and baseline system evaluation", 2017.

[ONE18-D31] ONE5G Deliverable D3.1. Preliminary multi-service performance optimization solutions for improved E2E performance, Apr. 2018

[PBM+13] K.I. Pedersen, S. Barbera, P-H. Michaelsen, C. Rosa, "Mobility Enhancements for LTE-Advanced Multilayer Networks with Inter-Site Carrier Aggregation", IEEE Communications Magazine, May 2013.

[POB+15] I. Petrut, M. Otesteanu, C. Balint and G. Budura, "HetNet handover performance analysis based on RSRP vs. RSRQ triggers," 2015 38th International Conference on Telecommunications and Signal Processing (TSP), Prague, 2015, pp. 232-235.

[PPS18] K. I. Pedersen, G. Pocovi, and J. Steiner, "Preemptive scheduling of latency critical traffic and its impact on mobile broadband performance," in Proc. VTC, Porto, pp. 1-6, 2018.

[PTS+18] Popovski, P., Trillingsgaard, K.F., Simeone, O. and Durisi, G., 2018. 5G Wireless Network Slicing for eMBB, URLLC, and mMTC: A Communication-Theoretic View. arXiv preprint arXiv:1804.05057.

[Put05] M. L. Puterman. Markov Decision Processes: Discrete Stochastic Dynamic Programming. John Wiley & Sons, 2005

[RKF+18] C. Rosa, M. Kuusela, F. Frederiksen, and K.I. Pedersen, “Standalone LTE in Unlicensed Spectrum: Radio Challenges, Solutions, and Performance of MulteFire”, to appear in IEEE Communications Magazine, October 2018.

[ROS11] Thomas Rückstiess, Christian Osendorfer and Patrick van der Smagt, “Sequential Feature Selection for Classification”, in AI 2011: Advances in Artificial Intelligence, Berlin, pp. 132-141, 2011.

[SB98] R. S. Sutton and A. G. Barto, Reinforcement Learning: An Introduction. Cambridge, Massachusetts. MIT Press, 1998.

[SNH+18] D. Sabella, N. Nikaein, A. Huang, J. Xhembulla, G. Malnati and S. Scarpina, "A Hierarchical MEC Architecture: Experimenting the RAVEN Use-Case," 2018 IEEE 87th Vehicular Technology Conference (VTC Spring), Porto, pp. 1-5, 2018

[SP+15] K. Smiljkovikj, P. Popovski, and L. Gavrilovska, “Analysis of the decoupled access for downlink and uplink in wireless heterogeneous networks,” IEEE Wireless Communications Letters, vol. 4, no. 2, pp. 173–176, April 2015

[SP+19] B. Soret, P. Popovski, K. Stern, “A queueing approach to the latency of decoupled UL/DL with flexible TDD and asymmetric services", *submitted to* IEEE Wireless Communications Letters, 2019, available at https://arxiv.org/abs/1904.02068.

[SW+18a] R. Schoeffauer, G. Wunder, "A Linear Algorithm for Reliable Predictive Network Control" 2018 IEEE GLOBECOM, 2018.

[SW+18b] R. Schoeffauer, G. Wunder, "Predictive Network Control and Throughput Sub-Optimality of MaxWeight”, EuCNC, 2018.

[SW+18c] S. Stefanatos, G. Wunder, “Performance Limits of Compressive Sensing Channel Estimation in Dense Cloud RAN”, ICC, 2018.

[TE92] L. Tassiulas, A. Ephremides, "Stability Properties of Constrained Queueing Systems and Scheduling Policies for Maximum Throughput in Multihop Radio Networks," IEEE Transactions on Automatic Control, Vol 37, Nr. 12, pp 1936-1948, 1992

[Texas13] Texas Instruments, “CC1101 Low-Power Sub-1GHz RF Transceiver,” http://www.ti.com/lit/ds/symlink/cc1101.pdf, datasheet, Nov 2013.

[VAR93] P. Varaiya, “Smart cars on smart roads: problems of control,” IEEE Trans. on Automatic Control, pp. 195-207, vol. 38, Feb. 1993.

[VML+18] I. Viering, H. Martikainen, A. Lobinger, and B. Wegmann “Zero-Zero Mobility: Intra-Frequency Handovers with Zero Interruption and Zero Failures” in IEEE Network Magazine, pp. 48-54, March/April 2018.

[WCA+17] H. Wang, S. Chen, M. Ai, & H. Xu, “Localized Mobility Management for 5G Ultra Dense Network,” IEEE Transactions on Vehicular Technology, vol. 66(9), pp. 8535–8552, 2017.

[WIN19] WINGS ICT Solutions “Serving underserved areas through 5G (IoT and Big Data) technologies”, https://www.youtube.com/watch?v=T97RdpSjbXg

[WMR10] Q. Wang, C. Mehlfuhrer, and M. Rupp, “Carrier frequency synchronization in the downlink of 3gpp LTE,” in 21st Annual IEEE International Symposium on Personal, Indoor and Mobile Radio Communications, pp. 939–944, Sept 2010

[WX06] X. Wang and W. Xiang, “An OFDM-TDMA/SA MAC protocol with QoS constraints for broadband wireless LANs,” ACM/Springer Wireless Networks, vol. 12, no. 2, pp. 159 – 170, 2006.

Acknowledgment

Contribution from the following colleagues is also acknowledged: Aspa Skalidi, Aikaterini Demesticha, Aimilia Bantouna, Apostolos Voulkidis, Vasilis Foteinos, Evangelia Tzifa, Ioannis Maistros, Konstantinos Tsoumanis, Kostas Trichias, Nelly Giannopoulou, Paraskevas Bourgos, Yiouli Kritikou, Ioannis Tzanettis, Georgios Makantasis, Despoina Meridou, Christos Ntogkas, Georgios Loukas, Vasilis Laskaridis, Ioanna Drigopoulou, Ilias Romas, Michail Mitrou, Kostas Tsagkaris (WINGS).

7 Appendix

7.1 Annex A RRC state handling for V2X URLLC

7.1.1 Road capacity

The formulation to determine road capacity is defined as:

$$C = v\frac{n}{ns + (n-1)d + D} \qquad \text{(A.1)}$$

where d represents the intra-platoon spacing, D is the inter-platoon spacing, s is the vehicle length, v is the steady-state speed, and n is the number of cars in each platoon.

7.1.2 Average latency

The average latency can be expressed as:

$$T = \sum_{N_V}\left[1-\left(\frac{N_s-1}{N_s}\right)^{n-1}\right]\Pr\left(N_v = n\right)\frac{nL_{pkt}R_{gen}}{S_{mcs}B} \le T_{\text{target}} \qquad \text{(A.2)}$$

where L_{pkt} is the size of the preamble, R_{gen} is the generation rate of random access preamble, S_{mcs} is the spectrum efficiency, N_s is the number of time slots for all vehicles and N_v is

$$N_v = \frac{BS_{mcs}\eta_{MAC}\eta_B}{L_{pkt}R_{gen}} \qquad \text{(A.3)}$$

Here η is the access efficiency, describing the level of access coordination.

7.1.3 Contention-based signaling efficiency

The contention-based signaling efficiency is defined as:

$$\eta_{MAC} = \frac{N_v}{N_s} \qquad \text{(A.4)}$$

where N_s is the number of slots to serve all users and can be expressed as:

$$N_s = \frac{S_{mcs}B}{L_{pkt}R_{gen}} \qquad \text{(A.5)}$$

7.1.4 Probability of collisions

The collision probability is defined as:

$$P_c = \sum_{N_V}\left[1-\left(\frac{N_s-1}{N_s}\right)^{n-1}\right]\Pr\left(N_v = n\right) \le \mathrm{P}_{\text{target}} \qquad \text{(A.6)}$$

7.2 Annex B Outage probability analysis with MC

In this appendix the outage probability gain and the resource utilization cost of MC with packet duplication are analytically derived.

7.2.1 System assumptions

A very flexible frame structure for 5G NR is introduced by 3GPP, offering different options to shorten the TTI duration compared to LTE [3GPP38.211]. We assume that a minislot consists in two OFDM symbols, corresponding to a TTI of 0.143 ms. This allows sufficient time budget for processing time, HARQ feedback and a single re-transmission within the one ms URLLC latency target. The metadata (i.e., control information) and the data are separate encoded with different BLER targets. This allows the metadata to be processed and decoded as soon as it is received, i.e. while data is still being received. From a latency perspective, this is advantageous as it allows performing the channel estimation earlier and enables early HARQ feedback as detailed in [BKP+16].

We assume that the meta data and the data are encoded separately with BLER targets respectively given by P_e^m and P_e^d. Note that, the outage probability of the data is reduced after chase combining following a possible retransmission. In general, we use $P_e^{d,l}$ to denote the outage probability after chase combining l copies of the packet; where $P_e^{d,l} < P_e^{d,k} \forall k < l$.

7.2.2 Baseline outage probability in single connectivity scenario

The possible events upon transmission are depicted in Figure 7-1.There are three different possible outcomes of processing the first transmission at the receiver: failure to decode the metadata (with probability P_e^m), metadata decoded, but failure to decode the data (with probability $P_e^d(1 - P_e^m)$), and successful decoding of the data packet (with probability $P_{succ}^{SC,1} = (1 - P_e^d)(1 - P_e^m)$.

In the event of receiving a HARQ NACK, the packet is retransmitted. If the metadata is successfully decoded, reception of the data is attempted by chase combining the retransmitted data with the initially received data. On the other hand, HARQ time out occurs if a HARQ feedback is not received within a pre-defined time interval. The transmitter then retransmits the packet automatically. However, there is no chase combining during decoding in this case, since the control information needed to identify the packet could not be decoded in the first instance. In this study, we set the HARQ time out to be equal to the round trip time, ensuring the same retransmission latency for both instances.

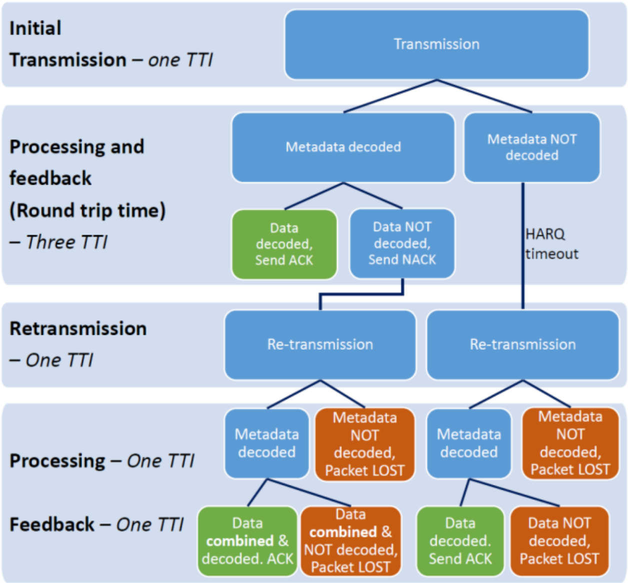


Figure 7-1 Schematic of events with single connectivity in the downlink direction

The success probability following retransmission is the sum of the success probabilities after blind retransmission (given by $P_e^m P_{succ}^{SC,1}$ and upon retransmission following a NACK. The latter can be either through direct decoding of the retransmitted packet, or upon combining with the earlier received packet. After some algebraic manipulation, the sum can be expressed as

$$\boldsymbol{P_{succ}^{SC,2} = P_e^m P_{succ}^{SC,1} + P_e^d (1 - P_e^m)^2 \big(1 - P_e^d P_e^{d,2}\big)}. \qquad \text{(B.2)}$$

Consequently, the total outage probability for the baseline single connectivity scenario is

$$P_{out}^{SC} = 1 - P_{succ}^{SC,1} - P_{succ}^{SC,2}. \qquad \text{(B.3)}$$

7.2.3 Outage probability in multi-connectivity scenario

Assuming independent transmissions of the same data packet over Mnodes, the packet is lost if it is not successfully decoded from any of the M nodes. Hence, the outage probability is given by

$$P_{out}^{MC}(M) = \prod_{m=1}^{M} P_{out,m}^{SC}, \qquad \text{(B.4)}$$

where $P_{out,m}^{SC}$ is the outage probability through the m^{th} node. In the case of identical outage probabilities through all links (i.e., with the same target BLERs), this becomes

$$P_{out}^{MC}(M) = (P_{out}^{SC})^M. \qquad \text{(B.5)}$$

7.2.4 Preliminary results and applications

The two different derived outage probabilities, namely that with SC and MC, are presented as a function of the BLER target on the data channel (P_e^d) in Figure 7-2 Outage Probabilities with SC and DC as a function of $\boldsymbol{P_e^d}$ for $\boldsymbol{P_e^m = 1}$%.. The BLER target for the meta data (P_e^m) is fixed at 1%, while the number of nodes is fixed at $M = 2$.

With SC, the outage probability remains above the URLLC target of 10^{-5} for P_e^d as low as 1% even after a single HARQ retransmission. However, such a target can be met with MC through two nodes. However, this is achieved at the expense of increased coordination and resource usage. In fact, since two nodes are used to transmit the same packet, the resource usage of MC is almost twice that of SC.

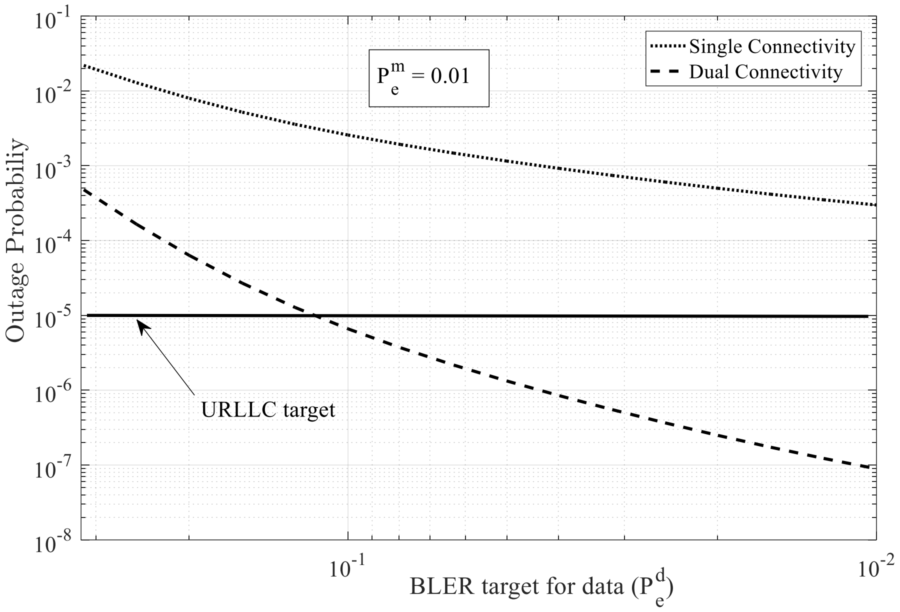


Figure 7-2 Outage Probabilities with SC and DC as a function of P_e^d for $P_e^m = 1\%$.

The derived outage probability equations can be used to determine the target BLERs required to achieve the URLLC outage probability target for a given transmission scheme. Figure 7-3 shows an example of such BLER target tuples required to achieve $1 - 10^{-5}$ reliability for SC and DC respectively. We observe that the desired URLLC reliability can be achieved with $P_e^m = 0.1\%$ and $P_e^d = 0.45\%$, whereas for DC, much higher BLER targets of $P_e^m = 2\%$ and $P_e^d = 7\%$ are required.

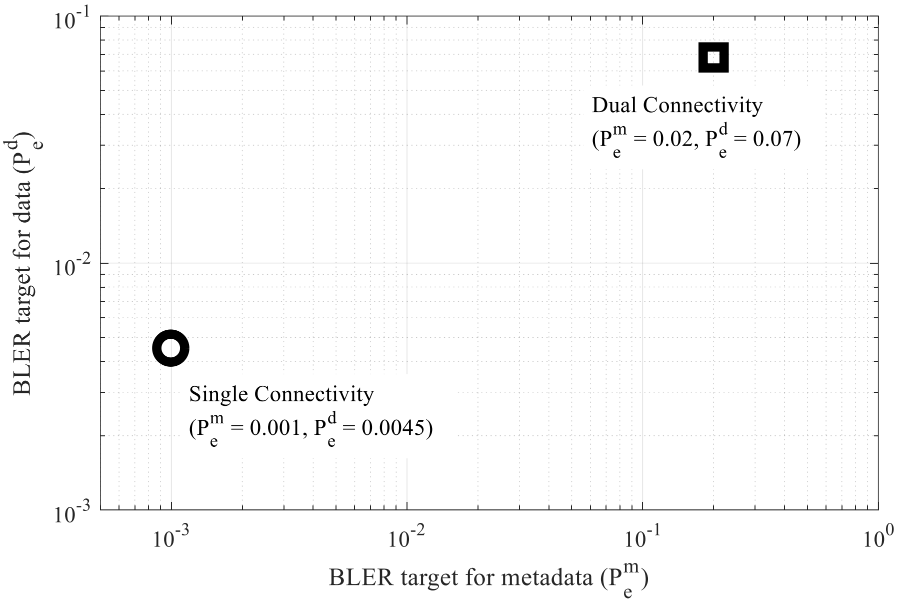


Figure 7-3 BLER target tuples for achieving $1 - 10^{-5}$ reliability with SC and DC

7.3 Annex C Baseline distributed scheduler (D-RAN)

The baseline distributed scheduler follows the same approach as explained in sub-section 3.1.4, except that it will be executed locally by each cell and with lack of neighbour cell information to efficiently manage harsh interferences. Moreover, CoMP and NOMA techniques will not be implemented, as no neighbour cell information will be available. Figure 7-4 and Figure 7-5 illustrate the performance of this algorithm under a distributed topology, denoted here as D-RAN Scheduler, compared to a traditional technique such as PF Scheduler. These figures show the results obtained by running different Monte-Carlo simulations under two different scenarios, Manhattan and Canonical, respectively, having the system configuration described in Annex E,Table 7-1.

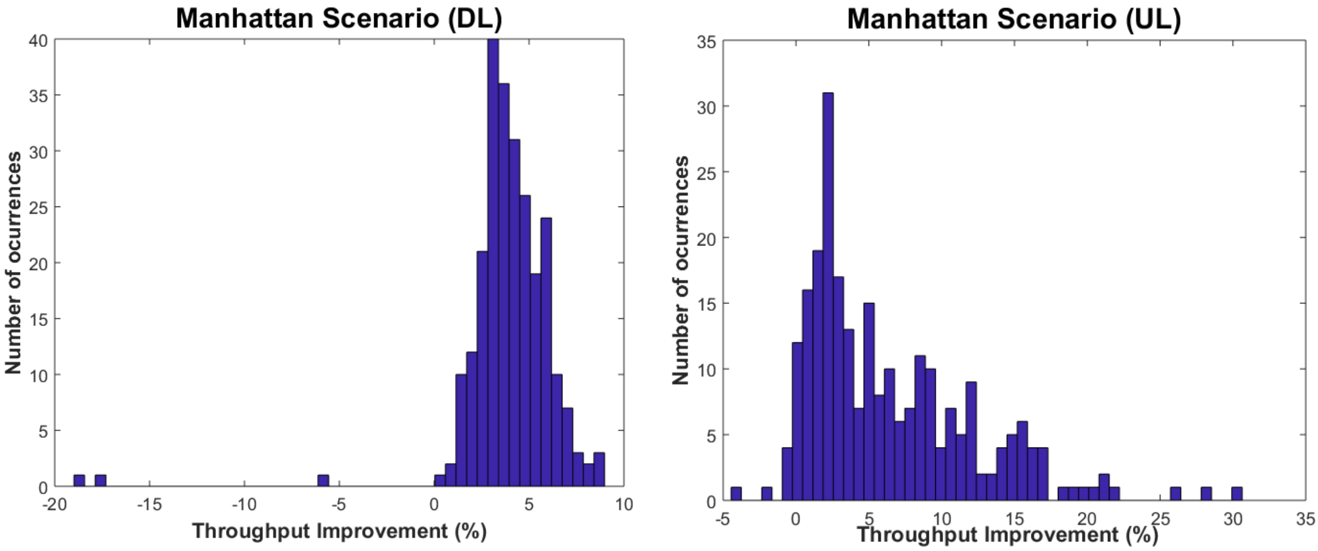


Figure 7-4 D-RAN Scheduler vs. PF D-RAN (Manhattan Scenario)

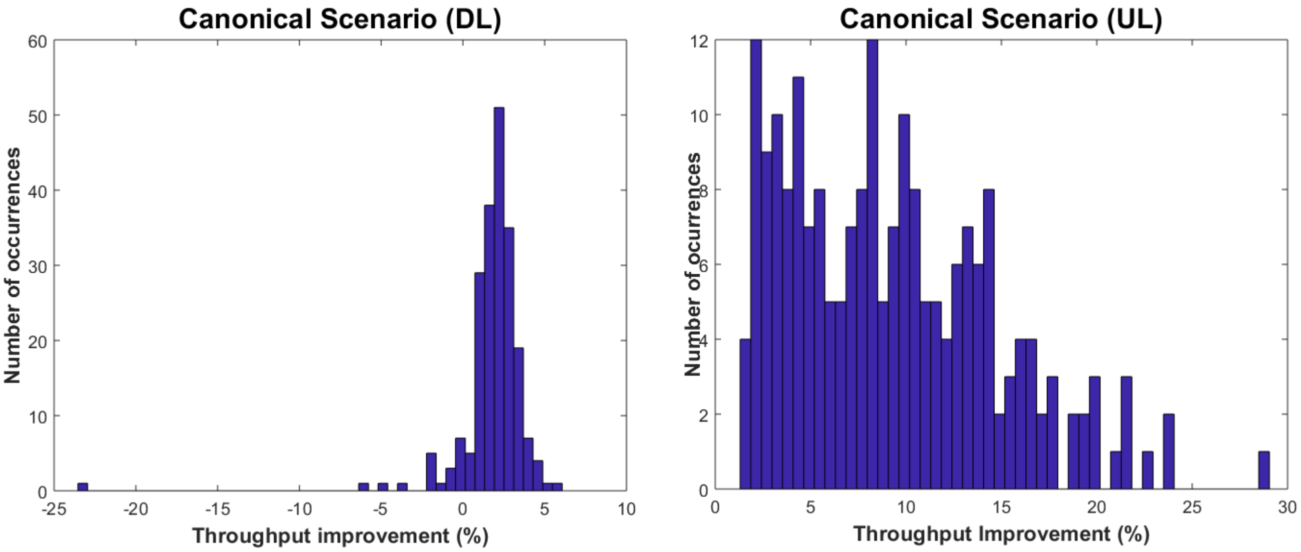


Figure 7-5 D-RAN Scheduler vs. PF Scheduler (Canonical Scenario)

As it can be seen in Figure 7-4 and Figure 7-5, the distributed scheduler (D-RAN) allows improving the system performance by boosting the aggregated throughput per cell. From the simulations depicted above, the aggregated throughput by implementing D-RAN scheduler increases roughly around 5%, in DL, and 5-15%, in UL, compared to the resulting aggregated throughput by using PF scheduler.

The improvement achieved by D-RAN scheduler is due to its ability on selecting the best parts of the bandwidth for users to be allocated based on their channel conditions and the possibility of jumping between the available subbands if better channel conditions are found. By means of this approach, harsh interferences affecting to overall system capacity can be softly mitigated, as D-RAN scheduler would select efficiently those bandwidth portions, which are less affected by neighbour cells/users interferences.

Moreover, there are few cases where there is no throughput improvement by using this scheduler. The reason is due to the suitability of the proposed scheduler for high-dense scenarios, where a large number of users are competing to get resources. PF scheduler performs user association just looking at the maximum metric in each subband, thus, PF will always select maximum user in each subband without looking at adjacent subbands. Conversely, the proposed distributed scheduler will look at adjacent subbands in order to select user-subband association less affected by interferences. But it may occur that for a few users, and with a maximum of subbands per user, distributed scheduler results in less throughput as adjacent subband checkability would lead to less throughtput. Let us give an example to better understanding this concept:

Taking into account following scheduling metrics to be used by both, PF and D-RAN schedulers and a maximum of 3 subbands:

Users	Subband 0	Subband 1	Subband 2	Subband 3	Subband 4	Subband 5
Imsi 1	10	3	5	2	20	6
Imsi 2	5	8	5	4	2	10

PF will result in the following association:

- Imsi 1: Subbands 0, 2, 4
- Imsi 2: Subbands 1, 3, 5
- Total metric: 57

Whereas, the distributed scheduler will provide:

- Imsi 1: Subbands 2,4
- Imsi 2: Subbands 0,1,3,5
- Total metric: 52

Hence, it can be concluded that the proposed scheduler is mainly suitable for megacities where massive users aimed to get network resources is foreseen.

In such a way and in view of the above results, a C-RAN deployment implementing the proposed centralized multi-cell scheduler seems to improve the overall system capacity in a greater extent, compared to a traditional distributed topology. The centralized multi-cell scheduler will allow to mitigate interferences and adverse effects of the scenario by bringing a new dimension into play, Remote Unit (RU), besides subbands, among which the centralized scheduler can jump in case better channel conditions need to be sought. Thus, harsh interferences caused by the high-density scenario (Megacities) would be managed more efficiently through the CU compared to a traditional distributed environment, which results, in turn, in an improvement of overall system throughput.

7.4 Annex D Centralized multi-cell scheduler

Table 7-1 Simulation Parameters

Parameter	Manhattan Scenario	Canonical Scenario
Number of sectors	25	21
Base station TX Power	23 dBm	42 dBm
ISD	90 m	500 m
Type of environment	Urban Outdoor	Urban Outdoor
Location of the BS	At street intersection (h = 3m)	Central cell surrounded by a ring of cells.
Number of UEs	300	300
UE TX Power	23 dBm	23 dBm
Location of UE	Random	Random
FDD/TDD	FDD	FDD
Frequency	2600 MHz (Band 7)	2600 MHz (Band 7)
Bandwidth	20 MHz	20 MHz
Traffic type	Full-buffer	Full-buffer
Number of MonteCarlo Simulations	10	10

7.5 Annex E Predictive Network Controller Scenario Configuration

7.5.1 Routing Matrix

The routing matrix R is a representation of the possible link to be activated by the binary control vector . For example, $r_{1,1} = -1$ and $r_{1,2} = 1$ imply a discrete time link, the packet is forwarded from q_t^0 to q_t^1 in an instant . For the case of the exposed scenario, Figure 4-25 (Left), q_t^0 to q_t^1 represent a link between the source MgNB_1 and SgNB_1.

$$R = \begin{pmatrix} -1 & -1 & 0 & 0 \\ 1 & 0 & -1 & 0 \\ 0 & 1 & 0 & -1 \\ 0 & 0 & 1 & 1 \end{pmatrix} \quad R = \begin{pmatrix} -1 & -1 & 0 & 0 \\ 1 & 0 & -1 & 0 \\ 0 & 1 & 0 & -1 \\ 0 & 0 & 1 & 1 \end{pmatrix}$$

$$R = \begin{pmatrix} -1 & -1 & 0 & 0 \\ 1 & 0 & -1 & 0 \\ 0 & 1 & 0 & -1 \\ 0 & 0 & 1 & 1 \end{pmatrix}$$

The matrix Q is a diagonal matrix to identify the packet destination. For the case of the exposed scenario in Figure 4-25 (Left), the destination weight is the UE's buffer, q_3.

$$Q = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

7.5.2 System Level Simulation Configuration

Environment	**3GPP-Umi 3gNBs, 9 cells** **200 meters inter-site distance**
CQI Report	Periodicity: 5ms, wideband CQI. Latency: ideal
Antenna Setup	BS:8 Tx, UE: 2Rx
eMBB user	10 users/cell
Traffic	TCP, interarrival: 10us, B=50B
UE Speed	30m/s
Cell Synchronicity	Ideal

MgNB Tx	46dBm
SgNB Tx	30dBm

7.6 Annex F Decoupled UL/DL access for flexible TDD and heterogeneous TTI requirements

5G supports flexible Time Division Duplex (TDD), which allows adaptation to the latency requirements. However, this flexibility is not sufficient to support heterogeneous latency requirements, in which different traffic instances have different switching requirements between UL and DL. This is visible in a traffic mix of enhanced mobile broadband (eMBB) and ultra-reliable low-latency communications (URLLC), like illustrated in Figure 7-6. The eMBB device requires long DL transmissions followed by short UL ACKs/NACKs, whereas an interactive process has a stringent latency requirement and sends short UL/DL packets continuously.

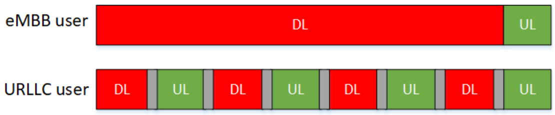


Figure 7-6 Traffic mix of eMBB and interactive URLLC

It is well known from queueing theory that waiting in a single line with two available servers is on average better than waiting in separated lines with one server each. The intuition behind is that tasks with a long task in front of the queue shall wait for a long time if only one server is available, whereas having a second server reduces the blocking situations. Translating this principle to a cellular system, **decoupling the UL and DL directions in a TDD network can be used for reducing the latency** (particularly of the URLLC devices) when having different switching requirements.

Decoupled access has been studied in the context of 4G Heterogeneous Networks (HetNets) to improve the average throughput. Having the focus on the user association and the interference, the related literature has used stochastic geometry for the analysis [SP+15]. In 4G TDD, the switching time limits the performance. With flexible TDD, the switching time is not a bottleneck anymore, because it is rather short especially in indoor scenarios. Another research question has been the possibility of having a link to more than one transmission point from the throughput and reliability perspective, but typically limited to one of the two transmission directions. The novelty of this proposal is the use of decoupled access for reducing the latency.

In [SP+19], the latency of decoupled access is analysed using queueing theory. The considered scenario is a TDD dense cell deployment with a central baseband pool connected with a fronthaul to a large number of small cells (RRHs). The system is abstracted in the queueing model in Figure 7-7, where the baseline coupled access is on the left, and the studied decoupled access in the right. The traffic is abstracted in two different queues: one for short TTIs and the other one for long TTIs. The short TTIs queue has strict priority over the long TTIs queue.

Naturally, the lack of coordination in the transmission directions leads to inter-RRH and inter-device interference. With the focus on the queueing gains, the details of the interference management are left for further study. In the results, it is assumed that the inter-RRH interference is ideally cancelled by sending the signal of the DL RRH to the UL RRH, such that it can be subtracted from the received signal. For the inter-device interference, a parametric approach maps the average interference level to a transmission latency. The channel is assumed to be Rayleigh. The eMBB device uses adaptive transmission rate, whereas the URLLC traffic has fixed transmission rate. URLLC traffic has priority over eMBB, meaning that if the long TTI queue is served only when the short TTI queue is empty.

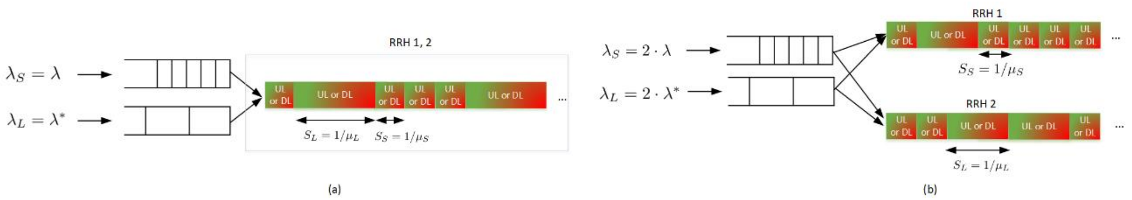


Figure 7-7 Queueing model with flexible TDD and devices with long and short TTI requirements. (a) Standard coupled access. Devices get the UL and DL from the same RRH. All the RRHs in the pool coordinate the transmission direction. (b) Decoupled access. Devices may receive the UL from one RRH and the DL from another one. The RRHs do not necessarily coordinate the transmission direction.

The results of the average sojourn time (queueing time + framing time + transmission time) are plotted in Figure 7-8. The short TTI is set to 1, and the long TTI takes the values 2, 10 or 15 depending on the channel quality (with two thresholds at 0 dB and 10 dB). The total sojourn time is plotted versus the system utilization. Naturally, devices with long TTI spend more time in the system, due to the longer service time and the low priority. Moreover, the decoupled access reduces the average time, and the improvement is remarkable as the intensity increases, corresponding to cases in which long tasks keep the server busy with a higher probability. The details of the analysis can be found in [SP+19].

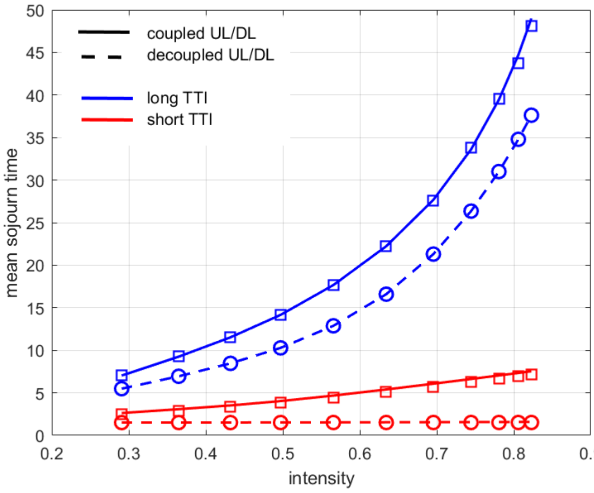


Figure 7-8 Results comparing coupled access and decoupled access

7.7 Annex G In band signaling optimization in LAA: CDRS

We use the indoor scenario proposed by 3GPP to evaluate coexistence with WiFi in unlicensed bands [3GPP36.889] to optimize the DRS in-band signaling. KQI integrity indicators i.e "File Transfer Average Throughput" service and "File transfer Delay" for FTP Model I have been analysed with a Poisson arrival distribution according to certain traffic intensity given by lambda λ (transfers per second) as an eMBB service representative.

In LAA, DRSs are used with two main functions: coarse synchronization and channel quality measurement. With regards to the first, DRS suppression would eventually lead to fewer coarse frequency and timing opportunities causing drifting/impairments in synchronization that could be compensated by Primary and Secondary Synchronization Signals (SS) inside the transmission bursts being channel estimation and detection its main function [WMR10]. Concerning this, we have observed a negative effect in MCS assignation procedure when DRS signals are disabled ("Slow Start"). Since no information regarding the quality of the channel is received, at the beginning of each LBT transmission, the assigned MCS index is the lowest. After 7 slots the UE receives an appropriate MCS index according to CQI feedback sent to the eNodeB through licensed uplink signaling channel. Although a lower modulation is more resistant to collisions, LAA initial transmissions will reduce the total throughput achieved by the system for eMBB FTP service Model I.

It is, therefore, essential to find a compromise between the service performance and in-band signaling functions preservation. As we exposed above, it would be potentially beneficial to increase DRS periodicity but still these opportunistic transmissions will cause a large number of collisions in the channel. In order to avoid this, suppression of DRS signals can overcome channel saturation but LAA performance is severely affected because of the already mentioned "Slow Start" MCS effect. Such effect produces lower data rates and increased delays. In short, we have improved fairness towards WiFi but at the cost of worsening LAA performance.

Therefore, it is necessary to find a compromise solution to enhance LAA performance reducing or eliminating the negative impact to WiFi when both technologies experience high traffic demand. The proposed solution to overcome the impact of suppressing DRS signals during dense coexistence aims to tackle directly the “Slow start" effect. To that end, a method called Compensation DRS (CDRS) covers the lack of channel estimation assigning the last stored MCS for that particular UE.

In that case, instead of starting the transmission with lowest MCS index, it will begin with an MCS matching the last perceived channel condition. Supposing a low mobility scenario, such assignment should not lead to high number of losses. However, in case of fast changing channel conditions, if the assigned MCS is above SNR level, some losses can be found.

Since DRS function is essential for good LAA performance, it is necessary to reactivate their opportunistic transmission when the channel conditions allow it. In this context, an adaptive omission of DRS, depending on load or number of successive times that LAA detects the channel busy, is suggested. Thus, when there are not many concurrent UEs in the scenario, DRS sending should be enabled back.

In the Figure 7-9 and Figure 7-10, we can see the results obtained for all possible DRS periodicity configurations, disabled DRS (NO DRS) case and our proposed compensation method (CDRS). In each case the WiFi KQI performance for different traffic intensities (lambda ranging from 0.5 to 2.5 values) is also shown.

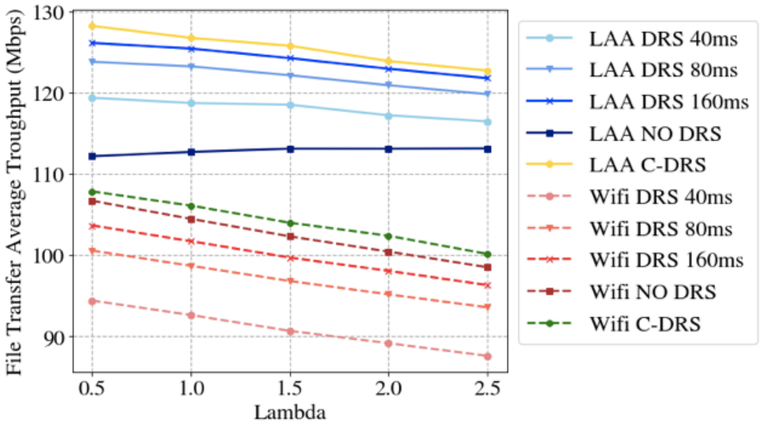


Figure 7-9 File Transfer Average Throughput evaluation

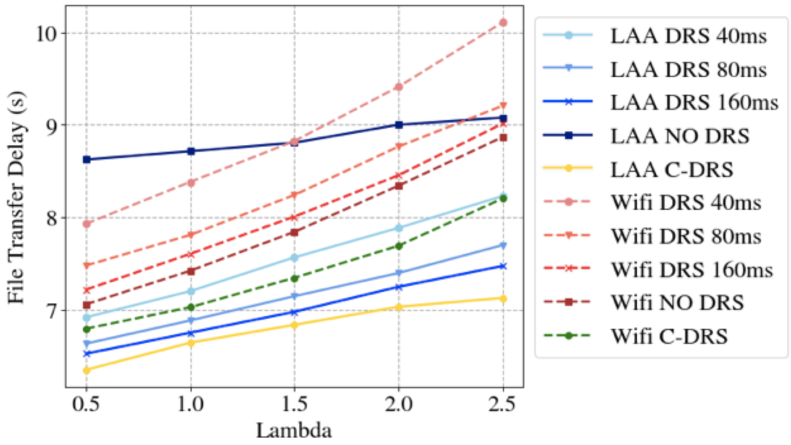


Figure 7-10 File Transfer Delay evaluation

As commented in previous paragraphs, due to “Slow Start” MCS assignment effect, LAA NO DRS case improves WiFi coexistence respect to any of DRS periodicity configuration cases (40, 80 and 160 ms). Nevertheless, as we predicted, the KQI values obtained are noticeably degraded. Finally, it can be verified that our proposed CDRS method obtains the best performance of all in terms of the end-to-end performance measured by File Transfer Delay and the File Transfer Average Throughput, while preserving the WiFi performance. In best case CDRS performs up to 40% better than best DRS standard configuration (160 ms)