Call:H2020-ICT-2016-2

Project reference: 760809

Project Name:

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

Deliverable D3.1
Preliminary Multi-Service Performance Optimization Solutions for Improved E2E Performance

Date of delivery:	30/04/2018	Version:	1.0
Start date of project:	01/06/2017	Duration:	24 months

Document properties:

Document Number:	D3.1
Document Title:	Preliminary Multi-Service Performance Optimization Solutions for Improved E2E Performance
Editor(s):	Mohamad Assaad
Authors:	Ana Herrera (UMA), Andreas Georgakopoulos (WINGS), Cesar Vargas Anamuro (ORANGE), Claudio Rosa (Nokia), Daniela Laselva (Nokia), David Palacios (UMA), Dong Min Kim (AAU), Eduardo Baena (UMA), Elena Serna Santiago (TID), Evangelos Kosmatos (WINGS), Francisco Javier Lorca Hernando (TID), Gerhard Wunder (FUB), Honglei Miao (Intel), Ioannis-Prodromos Belikaidis (WINGS), Isabel de-la-Bandera (UMA), Jens Steiner (Nokia), Jimmy Jessen Nielsen (AAU), Klaus Pedersen (Nokia), Koen de Turck (CNRS), María Luisa Marí (UMA), Matías Toril (UMA), Miltiadis Filippou (Intel), Mohamad Assaad (CNRS), Wenjie LI (CNRS), Juwendo Denis (CNRS), Nesrine Ben Khalifa (CNRS), Mustafa Emara (Intel), Nadège Varsier (ORANGE), Panagiotis Demestichas (WINGS), Raquel Barco (UMA), Richard Schöffauer (FUB), Rodolphe Legouable (BCOM), Saad Kriouile (CNRS), Salvador Luna (UMA), Sergio Fortes (UMA), Stelios Stefanatos (FUB), Vera Stavroulaki (WINGS), Yinan Qi (SEUK)
Contractual Date of Delivery:	30/04/2018
Dissemination level:	PU
Status:	Final
Version:	1.0
File Name:	ONE5G_D31_v10__final

Abstract

In this deliverable, we present our preliminary findings related to multi-service performance optimization solutions for improved E2E Performance. Our reference QoS architecture and protocol stack is first presented; including the PHY, MAC, RLC, PDCP, SDAP, and RRC layers in line with the 3GPP NR agreements. Moreover, different distributed and centralized RAN network architectures are described as will be used throughout the project. Preliminary work on different solutions to radio resource management for improving the E2E performance is presented. Among others, this includes different perspectives of resource allocation optimization for multiple services, signaling and control plane optimizations, and virtualization of mobile device capabilities for different architectures. Moreover E2E-optimized multi-link management techniques in the form of multi-node connectivity, dynamic spectrum aggregation mechanisms (licensed and unlicensed), advanced mobility and load balancing techniques and device-to-device (D2D) communications innovations are outlined. The presented material is work in progress, and hence will be further developed in the coming WP3 deliverable.

Keywords

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

Executive Summary

In this deliverable, we present our preliminary findings related to multi-service performance optimization solutions for improved E2E Performance. The presented material is work in progress, and hence does not present final findings from WP3.

First, we outline our assumed architecture and protocol stack reference models that will be used throughout the ONE5G project by the different work packages. The RAN user plane protocol stack is described, in line with the 3GPP NR agreements. It is described how the combined QoE manager (aka application layer scheduler) in the SDAP (Service Data Adaptation Protocol) layer and the faster radio-aware packet scheduler in the MAC-layer play an important role in controlling the E2E performance in terms of end-user performance, and thereby improve the Key Quality Indicators (KQIs). Moreover, different distributed and centralized RAN (D- and C-RAN) network architectures are described.

Preliminary work on different solutions to radio resource management for improving the E2E performance of 5G NR is presented. This includes the study of RRC state handling and discontinuous reception (DRX), novel resource allocation techniques that account for the different requirements and properties of different services. The former includes studies of signaling and control plane optimizations. Among others, this includes different perspectives of resource allocation optimization for multiple services. Solutions on signaling and control plane optimizations are presented as well, including studies of virtualization of mobile device capabilities in C-RAN and multi-access edge computing (MEC) for different architectures and functional split options.

Another primary focus of investigation is E2E-optimized multi-link management, including optimized usage of decoupled uplink and downlink cell associations, dynamic spectrum aggregation mechanisms, advanced mobility and load balancing techniques and device-to-device (D2D) communications. A mechanism to decide on the number of radio links and their usage (data split versus data duplication) for each user to fulfill certain QoS requirements is proposed, based on machine learning approaches. Research on dynamic spectrum aggregation is conducted from different directions. On the one hand, it is targeted through an evolution of MulteFire (the LTE standalone operation in the 5 GHz unlicensed band), targeting to reduce latency while improving reliability. On the other hand, joint usage of licensed and unlicensed bands is studied, including also ultra-dense urban environments.

Regarding mobility and load balancing enhancements, novel context-aware and QoE-aware traffic steering mechanisms are proposed. This includes cases also with MEC-aware UE-cell connectivity and a case of decoupled uplink/downlink connectivity. As for D2D, schemes for D2D association for interference reduction are outlined and also relay-based schemes for coverage enhancement and reduction of power consumption are studied.

Table of Contents

List of Figures

List of Tables

List of Acronyms and Abbreviations

3GPP	3rd Generation Partnership Project
ACK	ACKnowledgement
AP	Access Point
BBU	Baseband Unit
BWP	BandWidth Part
CA	Carrier Aggregation
CB	Code Block
CBG	Code Block Group
CC	Component Carrier
CCA	Clear Channel Assessment
CG	Cell Group
CN	Core Network
CoMP	Coordinated MultiPoint
COTS	Commercial Off-The-Shelf
CP	Control Plane
CPRI	Common Public Radio Interface
CQI	Channel Quality Indicator
C-RAN	Centralized RAN
CRE	Cell Range Extension
CRS	Cell specific Reference Signal
CSI	Channel State Information
CU	Centralized Unit
D2D	Device to Device
DCI	Downlink Control Information
DFT	Discrete Fourier Transform
DL	Down Link
DMTC	DRS measurement Timing Configuration
DPS	Dynamic Point Selection
D-RAN	Distributed RAN
DRB	Data Radio Bearer
DRS	Discovery Reference Signal
DRX	Discontinuous Reception
DVS	Device Virtualization Server
DU	Distributed Unit

E2E	End-to-End
eLAA	Enhanced LAA
eMBB	enhanced Mobile BroadBand
FDD	Frequency Duplex Division
FLC	Fuzzy Logic Controller
FTP	File Transfer Protocol
GBR	Guaranteed Bit Rate
gNB	5G NR base station node
HARQ	Hybrid Automatic Repeat Request
HetNet	Heterogeneous Network
HO	Handover
IoT	Internet of Things
IA	Interference Alignment
I-RNTI	UE-ID Radio Network Temporary Identifier
JT	Joint Transmission
JT/DPC-CoMP	Joint Transmission/ Dynamic Point Selection Coordinated Multipoint
KPI	Key Performance Indicator
KQI	Key Quality Indicator
LA	Link Adaptation
LAA	Licensed Assisted Access
LBT	Listen Before Talk
LTE	Long Term Evolution
LTE-U	LTE Unlicensed
LWA	LTE WLAN Aggregation
MAC	Medium Access Control
MC	Multi-Connectivity
MCS	Modulation and Coding Scheme
MEC	Multi-access Edge Computing
MF	MulteFire
MNC	Multi-Node Connectivity
MIMO	Multiple Input Multiple Output
mMTC	massive MTC
MTC	Machine Type Communication
MTD	Machine Type communication Device
NACK	Negative ACK
NAS	None Access Stratum
NFV	Network Function Virtualization

NGMN	Next Generation Mobile Network
NOMA	Non-Orthogonal Multiple Access
NR	New Radio
OAM	Operations, Administration and Management
OFDM	Orthogonal Frequency Division Multiplex
OFV	Objective Function Values
OLLA	Outer Loop Link Adaptation
PDCP	Packet Data Convergence Protocol
PDSCH	Physical Downlink Shared Channel
PF	Proportional Fair
PHY	Physical layer
PI	Pre-emptive Indication
PRB	Physical Resource Block
PSS	Primary Synchronization Signal
QoE	Quality-of-Experience
QoS	Quality-of-Service
RAN	Radio Access Network
RAT	Radio Access Technology
RLC	Radio Link Control
RNA	RAN Notification Area
RNAU	RAN Notification Area Update
RNIS	Radio Network Information Service
RRC	Radio Resource Control
RRH	Remote Radio Head
RRM	Radio Resource Management
RSRP	Reference Signal Received Power
RTT	Round Trip Time
Rx	Receiver
SCM	Spatial Channel Model
SDAP	Service Data Adaptation Protocol
SDL	Supplemental Down Link
SDN	Software-Defined Networks
SINR	Signal-to-Interference-plus-Noise Ratio
SSS	Secondary Synchronization Signal
SVD	Singular Value Decomposition
sRRT	Smoothened RRT
TB	Transport Block

TBS	TB Size
TCP	Transport Control Protocol
TDD	Time Duplex Division
TDMA	Time Division Multiple Access
TTI	Transmission Time Interval
UE	User Equipment
UL	Up Link
UP	User Plane
URLLC	Ultra-Reliable Low Latency Communications
SVD	Singular Value Decomposition
WDM	Wavelength-Division Multiplexing
WRC	World Radio Conference

1 Introduction

In this deliverable, we present our preliminary findings related to multi-service performance optimization solutions for improved end-to-end (E2E) Performance. Considering that this is the first deliverable originating from WP3, most of the materials are based on works that are currently under progress. Therefore, further findings with more details are yet to be provided in final WP3 deliverable.

In line with the ONE5G project proposal, we present solutions for optimizing the E2E performance of the 3GPP New Radio (NR). As a starting point, we explore detailed modeling of RAN related functionalities, which are the focus of this project. Namely, the relevant techniques in the project will operate at PHY/MAC/RLC/PDCP/SDAP and RRC layers including specific cross-layer functions. We explicitly consider realistic dynamics that occur during a data session, such as the transitions between the RRC/DRX states and effects of user mobility. Firstly, it is important to define *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 final user and reflects the behavior of the network and application as a whole, integrating the effects of various elements and configurations in the network over the E2E path, such as the RAN, core or external networks, interfaces, and the device/client itself. 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, abstract models of the impact with respect to latency and bandwidth originating in the fronthaul/backhaul network and the core are considered. Although further contributors to the E2E path are present in real networks, such as the processing at the client and a media server's capabilities to name a few, those elements are assumed to perform ideally for the reason that those are out of the scope in this project. In this deliverable, we describe *how E2E optimization can be achieved.*

E2E optimization is targeted either by designing enhancements and novel techniques or by configuring/tuning the values of the parameters relative to these techniques, such that in any case the investigated RAN-based techniques optimize selected indicators and metrics which, whenever possible, should reflect end-user performance. Particularly, in the context of ONE5G, we assume that optimizing the network towards improved E2E performance can be achieved by optimizing the Key Quality Indicators (KQIs) introduced in [D21]. Better KQI scores imply improvements of the performance phases associated to the given KQI categories, and in turn, provisioning of enhanced end user experience in term of objective Quality of Experience (QoE), which the KQIs are able to estimate. As explained in [D21], the KQI framework distinguishes five main distinct 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 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 integrating the KQI category scores properly 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), and we will assume that the KQI scores rely on either direct application-level KPIs or lower layer (network/radio) KPIs.

In Chapter 2, we present the considered architecture and protocol reference models in line with recent 3GPP NR decisions. Both the enhanced NR QoS architecture and protocol stack are described, as these constitute an important reference model for the project. In addition, the terminology used within the project is introduced. Furthermore, the various network options for distributed and centralized RAN realizations being considered are outlined, among others, building upon the reference cases being identified by NGMN. Finally, Chapter 2 outlines the control plane RRC machinery, the DRX scheme, and our assumptions for UE power consumption modeling being used throughout the project.

Chapter 3 captures several novel RRM enhancements for improving E2E performance. This includes the initial analysis of and the proposal of approaches for the management of the NR RRC machinery with IDLE, INACTIVE, and CONNECTED modes for various cases, including effects for DRX for UE power saving purposes. Secondly, a number of scheduler enhancements are studied. Those include pre-emptive scheduling, scheduler designs for cases with RAN slicing, MEC aware resource allocation, centralized scheduling solutions, and use of more advanced prediction techniques. Those scheduler enhancements all contribute to improve E2E performance. Finally, Chapter 3 also outlines methods for signaling and control plane optimizations, including device virtualization techniques and decoupled user- and control plane via different network nodes.

Chapter 4 addresses multi-link enhancements. Those include the following four main categories: (i) multi-link/multi-node connectivity for improved data rates and/or reliability, (ii) spectrum management and aggregation techniques, (iii) mobility and load balancing optimization, and (iv) performance optimizations for UEs using D2D links.

Finally, Chapter 5 concludes the report.

It is worth mentioning that, in collaboration with WP2, some contributions described in this deliverable will be implemented and validated in WP2 system level simulations. In particular, the techniques presented in Sections 3.2.5, 3.2.2, 4.1.3 and 4.3.1 are proposed for WP2 system level simulations.

2 Architecture and Reference Models

This chapter presents the relevant system reference models and the agreed terminology. In particular, the adopted QoS architecture, protocol stack, and the considered options for distributed- and centralized-RAN (D-RAN and C-RAN) implementations are presented. Finally, the control-plane protocol Radio Resource Control (RRC), Discontinuous RX (DRX), and related UE power consumptions models are introduced. This chapter therefore mainly outlines relevant information from 3GPP New Radio (NR) standardization and Next Generation Mobile Network (NGMN) as having relevance for ONE5G. However, while presenting this, we also add our interpretation and highlight related challenges and optimization opportunities relevant for the project.

2.1 QoS Architecture and RAN Protocol Stack

In ONE5G, we adopt the quality-of-service (QoS) architecture as agreed in 3GPP for the NR as our reference model throughout the project. This architecture was largely agreed during the NR Study Item, as captured in [38.804], and is now also captured in the NR stage-2 specification in [38.300]. This new QoS architecture offers multiple enhancements and degrees of freedom for enhanced multi-service orchestration, as studied in a recent paper [PPS+18], and thus also opportunities for more efficient KQI optimizations. How to best take advantage of this multitude of new options introduced with the enhanced NR QoS architecture and protocol stack is a challenging multi-dimensional problem that is yet to be solved and fully understood, and hence is one of the objectives of ONE5G. The 3GPP NR QoS architecture, which is depicted in Figure 2-1, has the following characteristics:

- For each UE, at least one E2E packet session is established (between UE and Peer).
- One or more QoS flows are associated to an E2E session.
- The 5G RAN associates the QoS flows with the data radio bearers (DRBs). This is conducted at the Service Data Adaptation Protocol (SDAP) [37.324] protocol layer in the 5G NR base station node (gNB), as will be explained in greater details later.
- This mapping is based on 5G QoS class indices (5QI) in the transport header of the packets, and on corresponding QoS parameters, which are signalled via the core network (CN) interface when a packet session is established.
- In the NR RAN, at least one default DRB is established for each UE when a new E2E packet session is created.
- On the 5G radio interface, the packet treatment is defined separately for each DRB. Different DRBs may be established for QoS flows requiring different packet forwarding treatment (e.g. associated with different requirements such as latency budget, packet loss rate tolerance, guaranteed bitrate - GBR) – e.g. QoS flow #1a and QoS flow #1b in the figure.
- On the terminal side, the concept of reflective QoS eliminates the need to use dedicated flow filters signalled by the network to match traffic to QoS flows. Thus, the UE derives the mapping of uplink traffic to QoS flows by correlating the corresponding downlink traffic and its attributes.

The 5QI (defined in [23.501]) contains a set of default QoS parameters for a large number of services, covering various enhanced mobile broadband (eMBB), ultra-reliable low latency (URLLC), and massive machine-type-communication (mMTC) use cases (see also ONE5G D2.1 [D21] for more information on use cases). The QoS parameters in the 5QI table include resource type (GBR, delay critical GBR, and non-GBR), priority, packet delay budget, packet error rate, and averaging window. Also, it is proposed in 3GPP (SA1 working group) to include information

such as expected packet inter arrival times and maximum packet size information for services where this is useful.

In line with [PPS+18], the default assumption in ONE5G is that all traffic for a UE is mapped to a single, or few, DRB(s), while the SDAP takes care of the differentiation; e.g. by modifying packet priorities, while only rarely modifying the QoS parameters of the DRB that the MAC-layer scheduler is to fulfill.

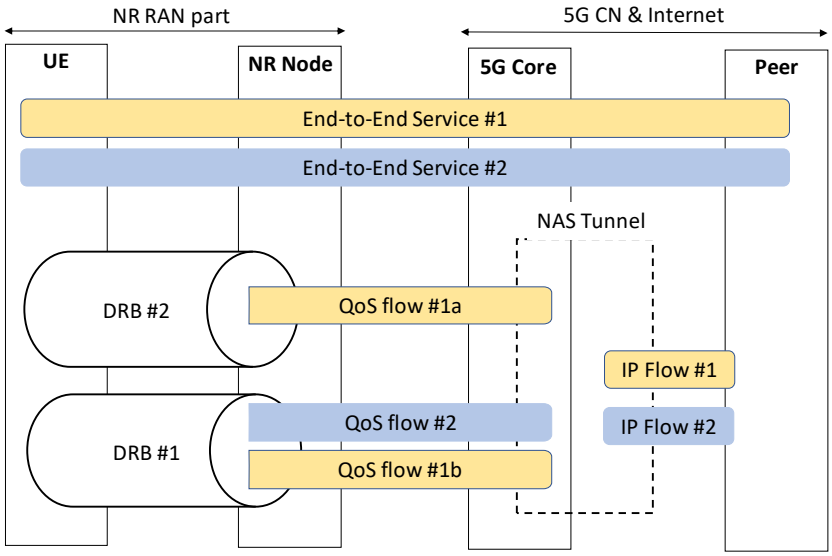


Figure 2-1: Overview of assumed QoS architecture in line with 3GPP NR agreements.

The user plane protocol stack in the NR RAN is summarized in Figure 2-2. Starting from the top, a new layer denoted by SDAP is introduced (compared to 4G). The SDAP is responsible for the mapping of QoS flows to DRBs, the marking of QoS flows, etc. Thereby, the so-called QoE Manager, can be implemented at the SDAP protocol layer [HSV16]. As an example, for the use case of YouTube streaming, the QoE manager may adaptively monitor and adjust the mapping of QoS flow(s) to the DRB, e.g. by adjusting the GBR and latency budget associated with the DRB to guide the MAC-layer radio scheduler. Thereby allowing to more efficiently optimize the KQIs as defined in [D21].

The NR packet data convergence protocol (PDCP) layer includes enhancements as compared to the PDCP for LTE. More importantly, the NR PDCP includes packet re-ordering (which for LTE was done at the RLC layer), as well as the possibility to enable packet duplication. Packet duplication is a powerful feature to enhance the reliability for UEs that are connected to multiple cells as is further studied in Section 4.1 as one of the available multi-connectivity options.

The NR Radio Link Control (RLC) layer is similar to its LTE equivalent, except from no longer having the packet re-ordering functionality as mentioned above. Secondly, RLC concatenation (as known from LTE) is replaced with MAC multiplexing for the NR, which offers further degrees of freedom for network implementations, where RLC and MAC are separated, and hence enables different C-RAN and D-RAN implementations [PPS+18].

The Medium Access Control (MAC) layer is the home of the radio packet scheduler and the Hybrid Automatic Repeat reQuest (HARQ) functionality, as well as the other functionalities listed in Figure 2-2. The MAC scheduler offers an extensive set of new scheduling options (e.g. different formats, variable TTI sizes, pre-emptive scheduling) that are further explored and studied in Section 3.2. In short, the MAC scheduler aims at fulfilling the QoS requirements for the DRBs. As we have identified in [PPS+18], the agile MAC scheduler should be carefully designed to work in harmony with the QoE management scheme at the SDAP layer to achieve optimized E2E performance.

The NR PHY layer includes a large set of enhancements and options for different configurations and flexibility. To mention a few, it includes options for configuration of different numerologies (e.g. subcarrier spacings), flexible frame structure, massive MIMO, etc. For more details, we refer to [38.802], [38.300] and [PPS+18].

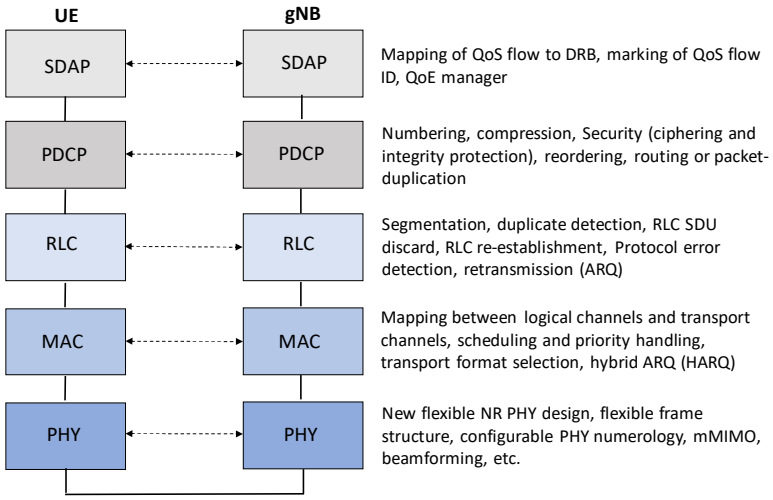


Figure 2-2: Overview of the assumed NR user-plane protocol stack.

The concept of network slicing is supported, as different traffic types may be handled by separate slices [MAB+16]. For example, this can be realized by mapping the data to different QoS flows/DRBs in coherence with slice-specific policies. On the device side, the NR supports UEs being able to provide information related to slice selection, provided by the non-access stratum (NAS) [38.804]. See more information on slicing in Section 2.2. The MAC scheduler simply aims at serving the users to fulfill the different DRB QoS requirements, and is in principle not aware of those relations to the same or different logical network slices.

2.2 D-RAN and C-RAN Architecture Options

Effective support of QoS parameters, as given in the 5QI table [23.501], largely relies on the ability of the network tenant to instantiate network functions at the most convenient location in the network. The concept of network slicing [NGMN15] can only show its full potential when it is connected to a higher degree of flexibility in the network topology. Assuming that the network can just be described as a collection of computing and storage nodes, provided with the proper connectivity between them, implementation of the different slices demands allocation of network function instances at different nodes, either closer to the user (e.g. for reduced latency) or deeper in the network (e.g. for better coordination). To enable this, C-RAN supports flexible functional location at RAN level.

The paradigm of C-RAN comprises two complementary approaches:

- Centralize those RAN functions that you can keep centralized *without impairing user plane performance*, and distribute only those that are required to be closer to the user.
- Virtualize the centralized RAN functions to get the advantages of Network Function Virtualization (NFV) and Software-Defined Networks (SDN) through extensive use of telco cloud technologies.

Centralizing RAN functions can be done in a varying manner, ranging from a totally distributed network a.k.a. D-RAN (Distributed RAN, as e.g. applied in LTE), to a fully centralized network or C-RAN (Centralized RAN). Any possibilities in between, leveraging telco clouds for the centralized elements, are under the scope of C-RAN.

C-RAN deployments traditionally leverage on the Common Public Radio Interface (CPRI) specification for feeding the fronthaul links between the remote sites and the Central Unit. CPRI poses stringent requirements in terms of bandwidth, latency and jitter to the transport network, especially when taking into account the bandwidths and bit rates that are foreseen in 5G. Hence, CPRI-based fronthaul links are generally based on dark fiber point-to-point connections, where the fibers are dedicated and point-to-point connections are not shared with any other transport service. In contrast, Ethernet infrastructure cannot meet the latency and jitter requirements of CPRI connection with IQ samples running at full rate irrespective of the actual traffic. Dark fiber deployments are generally very expensive, as they preclude the possibility to share the infrastructure with other transport services unless wavelength-division multiplexing (WDM) is employed, which requires additional active transport equipment.

In contrast, a more flexible fronthaul interface based on Ethernet has been the focus of C-RAN for 5G, with packetized IQ samples having less stringent requirements. The direct dependence of fronthaul bitrates with the actual cell traffic, which is not possible with CPRI, allows some data rate reduction through statistical multiplexing.

In addition, so-called non-standalone (NSA) deployments are the focus of the "early drop" of NR Release 15, where dual connectivity is the key lever to perform early NR deployments anchoring on legacy LTE networks. In this case, a device connects to a LTE eNB for basic coverage and control, and simultaneously to a 5G gNB as a secondary carrier component. Dual connectivity naturally relies on a given RAN split (at PDCP level) for anchoring the traffic flows to/from an LTE eNB and a 5G gNB (Figure 2-3).

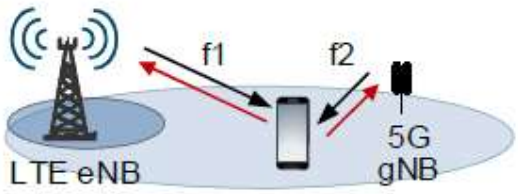


Figure 2-3: Schematic illustration of NSA mode in NR. f1 is the primary anchor carrier (LTE) and f2 is the secondary carrier (5G)

2.2.1 Agreed RAN architecture in 3GPP Release 15 (first drop of NR)

The agreed architecture in 3GPP follows a hybrid model (as shown in Figure 2-4). A 5G gNB can either be aggregated in a single node (left part in the figure), or disaggregated into three logical nodes (right part), namely: Remote Radio Head (RRH), Distributed Unit (DU) and Central Unit (CU). The latter can, in turn, be decomposed into CU-CP and CU-UP for control plane (CP) and user plane (UP), respectively. Proper interfaces are then further defined between these different logical RAN entities.

For simplicity, the split between DU and CU will follow one of only two possible options, namely:

Higher level split, agreed to be standardized between PDCP and RLC (Split 2, see

- Figure 2-5);

Lower level split, agreed to be at intra-PHY level but still pending discussions for future eventual standardization (Split 7, see

- Figure 2-5).

3GPP follows a pragmatic approach, where only the above two options are foreseen as the most interesting ones for 5G. Both splits lead to two new interfaces (F1 and F2), which in turn can be decomposed into the Control Plane and User Plane versions (F1-C, F1-U, F2-C, F2-U).

F1 is an IP/Ethernet interface (sometimes called "midhaul") that closely resembles the S1 interface of LTE, with not very stringent requirements in terms of bitrate, latency and jitter. F1 can easily benefit from statistical multiplexing because transport bitrates are proportional to the overall cell traffic, hence allowing a reduction in the transport capacity when aggregating multiple transport links.

F2 is, in contrast, an Ethernet-based interface carrying IQ samples (sometimes called "fronthaul") with stringent transport requirements, which was deferred to a later stage.

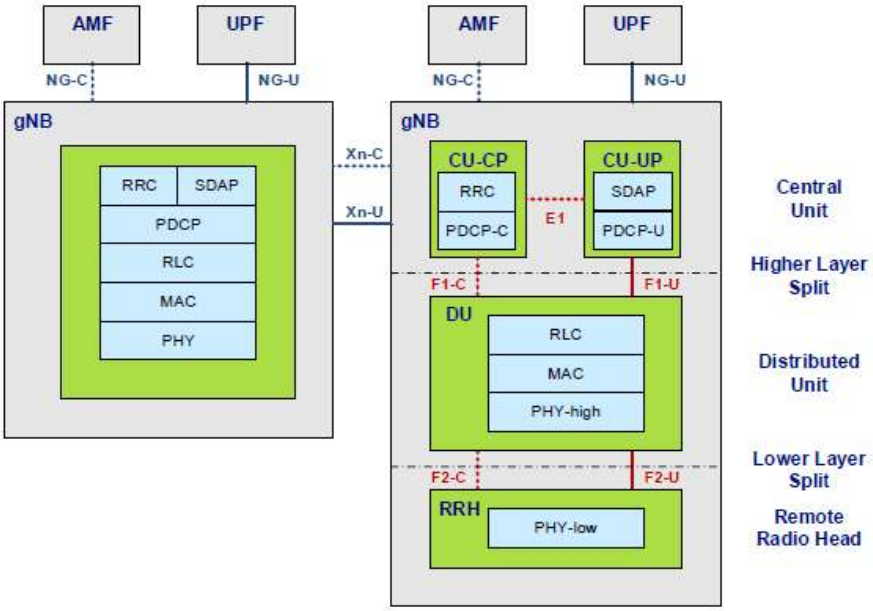


Figure 2-4: Schematic network architecture in 5G.

While the F1 interface will be part of NR Release 15, F2 will be implementation-dependent until the corresponding Study Item leads to a Working Item with concrete proposals for Release 16 or beyond. The most natural candidates for F2 are enhanced CPRI (eCPRI) and Next Generation Fronthaul Interface (NGFI), but there are other ongoing initiatives, like Extensible RAN (xRAN), Telecom Infra Project (TIP), etc. A common denominator for these activities is the reliance on Ethernet infrastructure, and the presence of significant bitrate compression compared to CPRI.

Figure 2-5 illustrates the options analyzed in 3GPP for the transport requirements in each case.

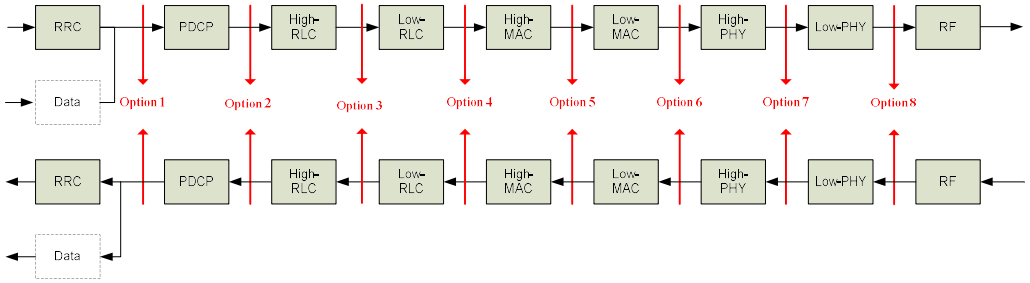


Figure 2-5: **Functional splits in 3GPP NR [38.804]**

The first eCPRI specifications [ECPRI17] have been released in August 2017 (V1.0). They offer a tenfold reduction in required transport throughput compared to CPRI and provide several split possibilities. Figure 2-6 illustrates the coexistence of non-IP eCPRI traffic and IP control traffic, together with basic synchronization and OAM (Operations, Administration and Management) means. Figure 2-7 illustrates some of the split options supported by eCPRI, although any other option is doable by means of proprietary signalling (out of the scope of the standard).

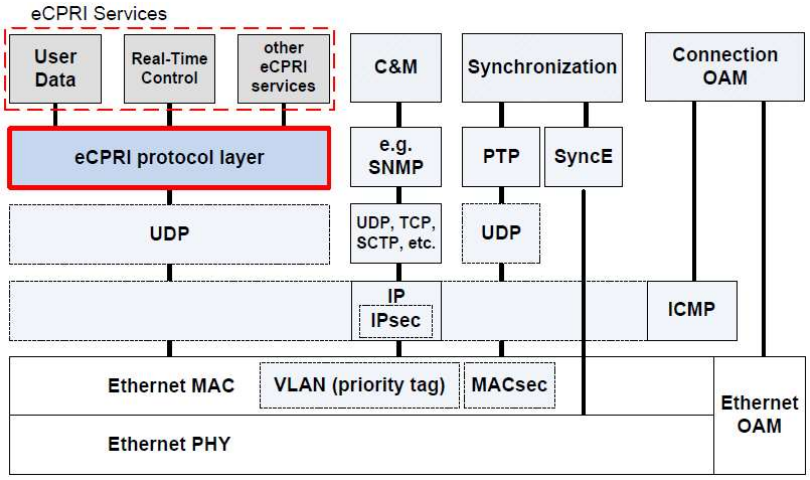


Figure 2-6: Functional diagram of eCPRI entities [ECPRI17]

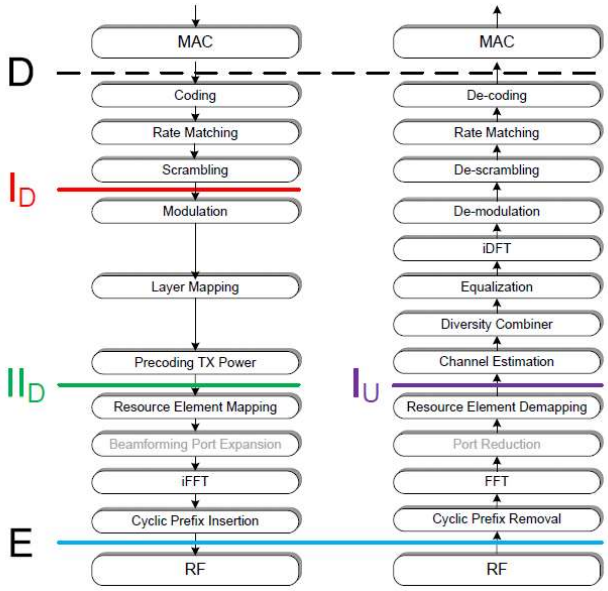


Figure 2-7: Split examples supported by eCPRI [ECPRI17]. Boxed represent specific tasks at Physical Layer level, and letters D, E, I_D, etc. refer to different Split options.

2.2.2 Agreed roles for the Central Unit and the Distributed Unit

With split 2 (as seen in

Figure 2-5, Option 2), the CU has the role of centralizing the control plane (RRM) and the PDCP sublayer in the data plane for a given number of cells. Hence, the CU performs RRM tasks and serves as an anchor node for NSA deployments. Given the high-level nature of those RAN tasks involved, the CU can leverage virtualization technologies running over Commercial Off-the-Shelf (COTS) hardware. However, PDCP encryption and integrity protection are processor-intensive tasks that could be accelerated with the aid of specialized cards (comprising GPUs, FPGAs or similar).

The DU performs the remaining data-plane functions, except the low-PHY that, in turn, resides at the RRH side, usually for a lower number of cells. The F2 interface is a logical interface, but actual implementations may integrate the RRH and the DU in a compact node, hence comprising all RAN functions from RLC to RF levels. For performance reasons, the DU is likely to run on dedicated hardware. However, the industry is reaching remarkable progress towards effective virtualization of the RAN lower layers, meaning that dedicated hardware implementations of the DU may soon be unnecessary.

The CU and DU allow varying levels of aggregation depending on the implementation needs. The CU could centralize e.g. up to several hundreds of sectors, given that RRM and PDCP layers do not present critical requirements, while the DU can aggregate e.g. several tens of sectors. Multi-access Edge Computing (MEC) can also be easily integrated with the CU for interesting applications related to e.g. video caching, local breakout, latency reduction or TCP optimization, among others.

2.2.3 Evolution towards network slicing

Network slicing is the most promising path to satisfy the requirements of the new services from vertical industries that can benefit from the well-known 5G main use case categories: eMBB, mMTC and URLLC. Besides the ability to create and/or eliminate slices (logical networks) dynamically, network slicing brings the potential to manage and orchestrate network resources in a centralized way, without the need to install, remove or re-configure the nodes in-site.

The network tenants can control their own slice via software, changing certain parameters to adapt the logical behaviour of the network to the requirements they want to fulfill at each moment. Moreover, the created slices are fully controlled by the operators, which can scale in/out network resources, i.e. dimension, according to the amount and type of traffic generated at a given point in time. An example of this could be an emergency situation (such as an earthquake, tsunami, etc.) where a slice devoted to emergency communications (by police, fireman, etc.) needs to be created or re-scaled with more stringent characteristics in terms of latency, compared to the other existing slices.

Network slicing logically transforms the network into a set of multiple independent networks, tailored to the specific needs of customers and running over a common physical infrastructure (Figure 2-8). The ambitious scope of network slicing leads to multiple Standards Groups extensively working on its definition and implementation, from 3GPP to NGMN, GSMA, ETSI, IETF or BBF.

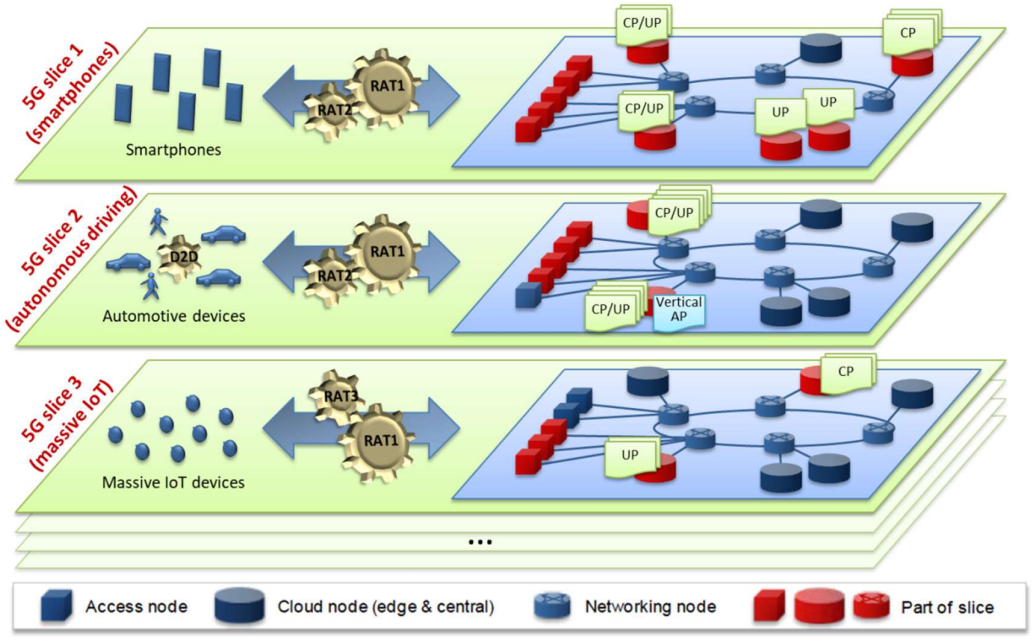


Figure 2-8: Network slicing architecture, as taken from [NGMN15].

As an essential pre-requisite to network slicing, the RAN nodes should be able to dynamically host different RAN functions as required by the different slices. This flexibility calls for C-RAN with ideally dynamic functional splits from high-layer to low-layer split, to address from very low latency services (with few RAN functions centralized and the use of MECs) to spectrally efficient advanced coordination techniques (with most RAN functions centralized).

2.3 RRC and DRX Reference Models

This section describes the Radio Resource Control (RRC) state machinery being related to the control-plane, as well as the Discontinuous Reception (DRX) framework along with the related UE power consumptions models, which are defined in 3GPP NR and will be adopted throughout our studies in ONE5G.

2.3.1 RRC state machinery

The RRC state machinery defines the procedures and the required signalling for each possible transition between different UE RRC states. The overall RRC state machinery for NR was agreed

as part of the NR study Item [38.804], as shown in Figure 2-9. [38.804] introduces a new independent RRC state, referred to as RRC INACTIVE, complementing the existing states, RRC CONNECTED and RRC IDLE, with the goal of lean signalling and energy efficient support of mMTC services. In fact, RRC INACTIVE allows mitigating signalling overhead and the associated state transition latency, therefore reducing network resource cost for small or sporadic traffic, while extending at the same time the UE battery life. Specifically, the RRC INACTIVE state enables to:

- Quickly start the transmission of small or sporadic data with low delay when the UE is in this RRC state;
- Minimize the required control signalling and the associated latency (comparably to the RRC CONNECTED state);
- Minimize the UE power consumption (achieving similar levels as with the RRC IDLE state); and
- Minimize network resource costs in the RAN/CN (achieving similar levels as with the RRC IDLE state) making it possible to maximize the number of UEs utilizing and benefiting from this state.

The related NR stage-2 specifications are part of [38.300] and, at the time of writing, the ongoing effort in 3GPP is to finalize the remaining Stage-3 signalling and procedural details for instance defining the number of and required RRC messages, and the information to be transmitted within a given message, for the Release-15 of the NR RRC specification [38.331].

Table 2-1 lists the features and characteristics that define the RRC INACTIVE state as compared to the RRC IDLE and RRC CONNECTED states. The RRC INACTIVE state is particularly beneficial for mMTC, i.e. for ONE5G use cases no. 3 and 4 (but naturally not limited to those) [D21].

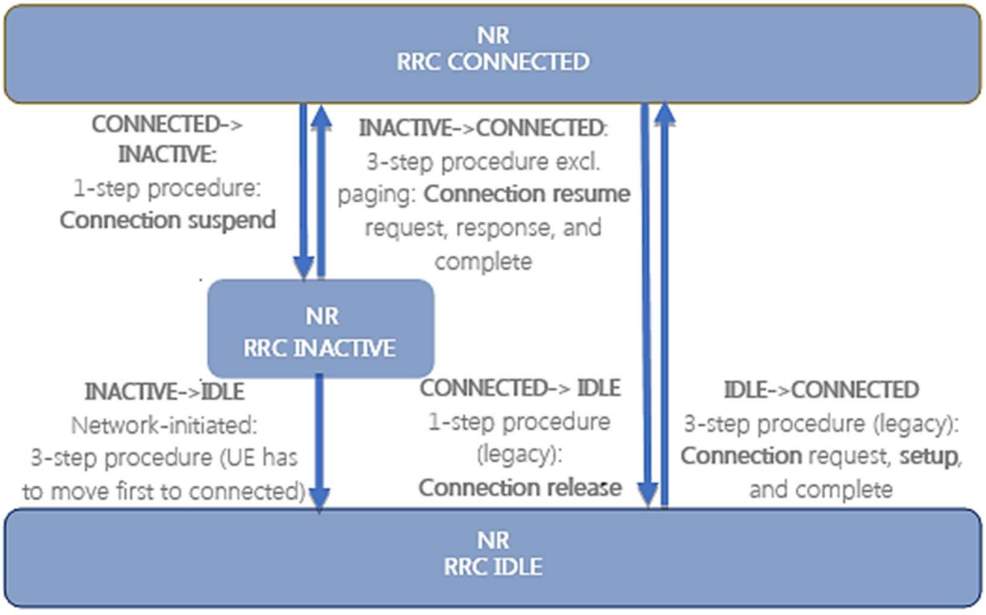


Figure 2-9: Overview of the RRC state machine and state transitions in NR.

Table 2-1: Comparison of RRC INACTIVE state characteristics vs. RRC IDLE and CONNECTED.

	No RRC connection has been established	**The RRC connection has been established**	
	UE in RRC IDLE	**UE in RRC INACTIVE**	**UE in RRC CONNECTED**
Established connections	None, i.e. no Control-Plane (CP)/User-Plane (UP) connections established.	The UE maintains CN/RAN connections (CP/UP) but no dedicated resources are reserved. The RAN knows the *cell area* the UE belongs to.	RRC connections are set up, including Carrier Aggregation (CA)/ Dual Connectivity (DC); The RAN knows the *cell* the UE belongs to.
Data transfer	Not supported; requires state transition	Small UL data supported Large UL data requires state transition	Data transfer data to/from UE fully supported
UE AS context	Not stored in gNB/UE	Stored in at least one gNB in the cell area and at the UE	Stored in NR RAN
DRX	Can be UE specific DRX	Can be UE specific DRX	Can be UE specific DRX
Paging	Paging initiated by CN Paging area managed by CN UE monitors paging channel	Paging initiated by NR RAN; *RAN-based notification area* (RNA) managed by RAN (incl. RNA location update when a UE moves out of the assigned area, i.e. RNAU)	Not needed
Measurements	UE performs neighboring cell measurements	UE performs neighboring cell measurements	UE performs neighboring cell measurements/reports UE monitors PDCCH control channel UE provides CQI/CSI/other feedback
UE mobility	UE-controlled, i.e. cell re-selection mobility	Cell re-selection mobility within RAN area	Network controlled, i.e. handover
System information (SI)	UE acquires “Minimum SI” (broadcasted periodically and always present)		
	UE can request “other SI” w/o a state transition	UE can request “other SI” w/o a state transition	Dedicated signaling for “other SI” request / delivery

The key aspect of the new state is the capability of the UE of transferring data without requiring prior transfer to the RRC CONNECTED state. For the sake of signalling reduction, the UE context (e.g. NAS/AS connection setup, bearer setup) which specifically includes the security context (i.e. PDCP encryption keys), is stored in at least one gNB in the assigned RAN area as well as at the UE. On one side, this removes signalling for UE context creation at the CN/RAN, however on the other side, security becomes challenging when a UE in RRC INACTIVE accesses a gNB other than the last serving gNB. For instance, at the connection resume the encryption keys will have to be changed if the UE resumes to a different gNB than the one storing the UE context (defined as the anchor gNB). For that purpose, the target gNB has to first retrieve the original UE context from the anchor gNB for a valid UE. A UE is valid if it can identify itself successfully to the RAN by the unique UE ID provided by the anchor gNB, known as resume ID, in the form of I-RNTI (UE-ID Radio Network Temporary Identifier). Upon successful context retrieval (via a XnAP Retrieve UE Context procedure), the target gNB becomes the serving cell and the UE context at

the anchor gNB can be released. If a valid UE Context cannot be found, a new RRC connection has to be established rather than resuming the previous RRC connection.

The mechanisms and message flow required for enabling small UL data transmission of a UE in RRC INACTIVE are yet to be finalized in 3GPP, and at the time of writing such support is expected to be part of Release 16. In this report, we assume as working assumption that the data transfer for UEs being in RRC INACTIVE is always preceded by a Random Access Channel (RACH) procedure according to [38.804]. That is, a contention based scheduling procedure is used with a two-step procedure where the UE requests for a grant may be granted if no collision happens. Upon receiving a successful random-access response, the UE would then indicate in the connection resume request its resume ID and consequently start transmitting the first UL data packet. Based on those assumptions, Table 2-2 lists for each RRC state the number of signaling messages required before the UE can start transferring the first uplink data packet. The table also indicates potential state transition options after the data transfer is complete. Rules for identifying when to use RRC INACTIVE state as well as when to move away from this state are not in the scope of 3GPP and will be investigated in the later chapter as part of the RRC states optimizations.

Table 2-2: ONE5G assumptions on transferring first UL data packet and state transitions after the data transfer.

	UE in RRC IDLE	**UE in RRC INACTIVE**	**UE in RRC CONNECTED**
Data transfer	*The UE requires 10 messages to transfer first UL data packet.* Note: legacy case, see [Fan5G D4.2])	*The UE requires 3 to 4 messages to transfer first UL data packet.* Note: assumes 2-step contention procedure.	*The UE requires 1 to 2 messages to transfer first UL data packet.* Note: legacy case, depending on whether the UE has already received the grant or not.
State transition after data transfer	-	After small data transfer the UE may move to CONNECTED or IDLE, or remain in INACTIVE. If large data transfer the UE will move to CONNECTED.	The UE moves to INACTIVE/IDLE after DRX inactivity timer(s) expire. Otherwise, the UE remains in CONNECTED.

2.3.2 DRX framework

In 3GPP NR system, to optimize the UE power consumption, discontinuous reception (DRX) is supported to dynamically switch the UE transceiver on-off according to the actual traffic demand. To realize DRX operation, the MAC entity may be configured by RRC with a DRX functionality that controls the UE's PDCCH monitoring periodicity and format. When in RRC_CONNECTED, if DRX is configured, the MAC entity may monitor the PDCCH discontinuously using the DRX operation; otherwise, the MAC entity shall monitor the PDCCH continuously. According to [38.331], RRC can control DRX operation by configuring the following timers, and the NR unit below refer to the scheduling time unit:

- drx-onDurationTimer: the number of consecutive NR unit(s) at the beginning of a DRX Cycle. Unit in milliseconds;
- 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;
- drx-RetransmissionTimerDL (per DL HARQ process): the maximum number of consecutive NR unit(s) until a DL retransmission is received;
- drx-RetransmissionTimerUL (per UL HARQ process): the maximum number of consecutive NR unit(s) until a grant for UL retransmission is received;
- drx-LongCycle: Long DRX cycle. Unit in milliseconds;
- drx-ShortCycle: Short DRX cycle. Unit in milliseconds;

- drx-ShortCycleTimer: the number of consecutive NR unit(s) the UE shall follow the Short DRX cycle;
- drx-HARQ-RTT-TimerDL (per DL HARQ process): the minimum amount of NR unit(s) before a DL assignment for HARQ retransmission is expected by the MAC entity;
- drx-HARQ-RTT-TimerUL (per UL HARQ process): the minimum amount of NR unit(s) before a UL HARQ retransmission grant is expected by the MAC entity.

Most timers in LTE MAC are configured by using the time unit of a PDCCH-subframe (psf). As there is greater flexibility in NR PDCCH configuration, the term PDCCH-subframe could cause some confusion in NR. For instance, if the UE is monitoring several component carriers using different numerologies, the number of PDCCH occasions and thereby PDCCH-subframes would be very different on different carriers and could be ambiguous in the framework of connected mode DRX. It is therefore agreed in NR that the timers drx-onDurationTimer, drx-InactivityTimer, drx-LongCycle and drx-ShortCycle are configured by using a time unit of milliseconds instead of PDCCH subframes (psf), as in LTE. However, the values for these timers in LTE with the unit in psf, could be easily translated into ms, resulting in the same operation options.

The drx-onDurationTimer in LTE is typically configured in the order of psf20, i.e. around 20ms. The reachability for the UE and the scheduling flexibility at the gNB depends on the length of the drx-onDurationTimer. Allowing durations of fractions of a ms may lead the UE to enter and leave DRX at non-slot boundaries, hence increasing the risk of state mismatch. The situation is similar for the drx-InactivityTimer. It aims to have the UE reachable a certain time after the last initial transmission of a transport block. This is to prevent the UE from entering sleep mode when the probability for new arrival of data is high. Having values less than a millisecond would also for this timer impose a risk of having the UE in an undefined state between active and sleep mode and increase the risk of state mismatch.

According to [38.211], a carrier bandwidth part (BWP) is a contiguous set of physical resource blocks selected from a contiguous subset of the common resource blocks for a given numerology on a given carrier. A UE can be configured with up to four carrier bandwidth parts in the downlink (uplink) with a single downlink (uplink) carrier bandwidth part being active at a given time. Moreover, a new timer (BWP inactivity timer), i.e., bwp-InactivityTimer, is also supported [38.331] to switch active BWP to default BWP after a certain inactive time. BWP inactivity timer is independent from the above DRX timers. For paired (unpaired) spectrum, a UE starts the timer when it switches its active DL BWP (DL/UL BWP pair) to a DL BWP (DL/UL BWP pair) other than the default DL BWP (DL/UL BWP pair). A UE restarts the timer to the initial value when it successfully decodes a Downlink Control Information (DCI) to schedule PDSCH(s) in its active DL BWP (DL/UL BWP pair). A UE switches its active DL BWP (DL/UL BWP pair) to the default DL BWP (DL/UL BWP pair) when the timer expires.

2.3.3 UE power modeling

UE power consumption models are useful for investigating the power consumption impact of different RRM requirements. The power consumption models are simplified models and are not intended to capture details or in any way limit UE implementation. In [38.803], four different power consumption states are defined as shown in Figure 2-10:

- Deep sleep: The UE is operating 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.
- Light sleep: The UE in this state has maintained timing using a clock and activity level which allows reception to be started with a reasonably small delay. This state represents the UE being ready to start to receive with minimal delay.

- Active RX only: The UE is actively receiving, or attempting to receive a signal. The RF receiver circuit is active in this state.
- Active TX only: The UE is actively transmitting a signal.

For active TX + RX where the UE is actively receiving and transmitting (RX and TX are active), the power consumption in this state is assumed to be the sum of the active RX only power consumption + active TX only power consumption.

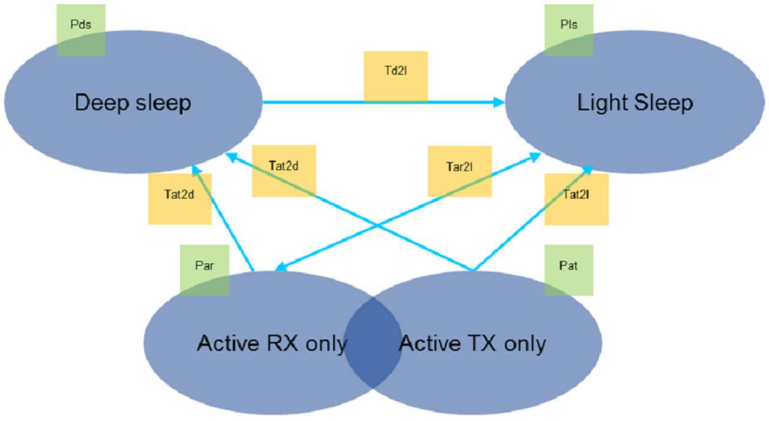


Figure 2-10: UE power consumption states model [38.803].

The UE power consumption state can change dynamically, based on NR UE requirements for reception and transmission. For example, if DRX is active in a UE it can be expected to spend as much of the DRX off duration as possible in deep sleep. Prior to an RX-on period, it may move to light sleep state before it enters active RX and TX, depending on requirements for reception and transmission. UE will remain in active state as required, according to monitoring and scheduling requests. When the UE is no longer required to receive or transmit, it may move back to light or deep sleep again. Power consumption may be estimated by considering the proportion of time that the UE spends in each state. RRM measurement and report activities may extend the duration that the UE needs to remain in either the active RX state or the active TX state.

According to [38.803], the parameters used for the NR power consumption model in Figure 2-10 are given in the following table.

Table 2-3: Predicted power consumption in different states for the UE model [38.803]

Parameter	Label/value	Default value
Absolute power consumption in "Active with data TX" state. Relative upper limit set to get some absolute numbers on the graphs	$P_{at, CA\text{-}N}$ Note 1	800 mW, [GTI17]
Relative power consumption in "Active data/no data RX" state	$P_{ar, CA\text{-}M}$ Note 2	600 mW, [GTI17]
Relative power consumption in "Deep sleep" state	P_{ds}	0.03 W, [LBS+14]
Relative power consumption in "Light sleep" state	P_{ls}	0.6 W, [LBS+14]
Duration of transition from "Deep sleep" to "Light sleep" state	T_{d2l}	20 ms, [GTI17]
Relative power consumption while changing from "Deep sleep" to "Light sleep" state	P_{d2l}	0.3 W [GTI17]
Duration of transition from "Light sleep" to "Active with (no) data RX" state	T_{ar2l}	20 ms, [GTI17]
Duration of transition from any active state to any sleep state	T_{at2d} or T_{at2l} or T_{ar2d} or T_{ar2l}	40 ms, [GTI17]
Note 1: This parameter is determined for each possible UL CA configuration. Note 2: This parameter is determined for each possible DL CA configuration.		

3 RRM for Improved E2E Performance

This chapter presents the work concerning the topic of Radio-Resource Management (RRM) techniques for improving the End-to-End (E2E) performance in 5G NR systems, or equivalently improving the defined KQI counters in [D21] for different service categories. The contributions on this topic are split into three categories. Initially, we present the work on optimized RRC state handling and DRX, where in particular the use of the new RRC INACTIVE state is considered, as well as the integration with network slices. Second, the contributions studying resource allocation and network slicing techniques are presented. Finally, the chapter presents work on signalling and control plane optimizations.

3.1 Optimized RRC State Handling and DRX

This section summarizes the ongoing studies oriented to optimize RRC and DRX operations and the associated new degrees of freedom enabled by the 5G NR design. In the following we will discuss how to optimize the RRC state transitions on a per service basis or for distinct network slices as well as how to optimize the DRX handling for bandwidth adaptation.

3.1.1 Optimized RRC state handling for NR

In this section, we discuss the RRC state handling for 5G New Radio deployments introducing the key factors that need to be considered when optimizing the RRC state transitions. The analysis in the following will be applicable to any service category, and therefore to any of the use cases developed in ONE5G. However, it will be particularly beneficial for those use cases targeting mMTC services, such as no. 3, 4, 7, and 8 [D21]. Furthermore, a dedicated application could be envisioned for the underserved areas where the power availability is expected to be limited.

As presented in Section 2.3, the RRC state machine designed for NR introduces the novel state of RRC INACTIVE, which is designed primarily to enable an optimized support of massive amounts of mMTC devices. It achieves that thanks to the reduction of control-plane signalling, while balancing between battery efficiency and low latency when accessing the network (i.e. control-plane latency). As the new RRC state introduces new degrees of freedom in how to operate the RRC state transitions, it brings further complexity to determine when to use each of the three RRC states beneficially. In this report, we'll focus on the RRC INACTIVE state and describe how the traffic characteristics (e.g. payload size, message arrival frequency) and service requirements such as the latency target, will impact its usage, highlighting the benefits when using it, and finally hinting on how to design an optimized RRC state handling framework capable to well balance the power consumption vs. latency trade-off.

To explain how closely the RRC state handling depends on the traffic characteristics as well as the benefits provided by RRC INACTIVE, we provide a couple of examples. As mMTC services may have typically low activity state, such that the device will wake up sporadically to transmit a small message, they need not use power-hungry RRC connected mode to transmit (or receive) the small payload, but rather use RRC INACTIVE to meet their stringent battery consumption limitation. Particularly, in case of periodic or frequent traffic demand, the use of RRC INACTIVE could avoid frequent RRC IDLE to RRC CONNECTED mode state transitions which consume control-plane signalling in the network, posing challenges to the maximum number of RRC CONNECTED UEs that can be supported simultaneously. Another example relates to URLLC services that demand ultra-low latency in accessing the network prior to the transmission of their small packets. Inherently, devices demanding URLLC services can benefit from RRC INACTIVE rather than RRC IDLE to guarantee the low access delay.

On a general level, it seems sensible that whenever there is no or limited activity from a UE during a rather short time, the network can suspend an ongoing RRC session reducing the amount of control-plane network resources assigned to the UE as well as limiting the UE power consumption, by moving the UE to RRC INACTIVE. Afterwards the network could resume the UE session moving the UE back to RRC connected mode. Thus, the high-level framework comprises determining for which combinations of service type, service requirements (latency, energy), mobility pattern, traffic profile and activity state, the following decisions should be taken:

For a UE in RRC CONNECTED:

- *With data activity:* the UE RRC connection should be suspended moving the UE to RRC INACTIVE, or rather retain the connection remaining in CONNECTED mode.
- *With no data activity:* the UE should be moved to RRC IDLE or suspend its connection moving it to INACTIVE, or rather remaining in CONNECTED mode.

For a UE in RRC INACTIVE:

- *With data activity:* the UE should transmit and receive data while in RRC INACTIVE or rather should resume its connection in CONNECTED mode.
- *With no data activity:* the UE should be moved to RRC IDLE or rather be retained in INACTIVE mode.

For a UE in RRC IDLE:

- *With data activity:* The UE should be moved to RRC CONNECTED to accomplish any data transmission.
- *With no data activity:* The UE can likely and safely remain in RRC IDLE until any data activity is requested.

In the following we elaborate in more details on the traffic properties, service requirements, and UE context properties which are relevant for the framework design, hinting to how each of these factors would influence the RRC state transition policy.

Service requirements:

QoS targets: Among the QoS targets described in the earlier sections, delay matters particularly in this context. For instance, the configured RRC state and state transitions should tightly depend on the user-plane latency target to make sure that the introduced network access latency (i.e. control-plane latency) prior to start any data transmission will not make the latency target unfeasible.

Power availability: Whenever limitations apply in the power availability, the RRC state transition policy should be carefully designed to allow the UE to move as soon as possible to an energy-efficient state (e.g. RRC INACTIVE or IDLE). The following power properties are defined in [22.368]:

- *External power supply available*
- *Rechargeable battery*
- *Non-replaceable battery*
- *Unknown*

Traffic priority:

- *Expendable/best-effort*: The traffic may be dropped during times of congestion due to level of importance, retransmission mechanisms, or error correction.
- *Pre-emptive*: The traffic may be dropped for critical traffic during times of congestion.

- *Critical*: traffic should not be dropped during times of congestion.

Traffic characteristics associated to the service:

Traffic profile: The classification of the traffic arrival characteristics is critical for an optimal selection of the RRC state transition strategy. As an example, although, it is sensible that a UE is moved to RRC IDLE after a sufficient long inactivity time, it should be avoided that the UE is moved to RRC IDLE shortly before it seeks to transmit new data. Optimizations are however possible whenever it appears some sort of temporal structure in the traffic. The main categories of the traffic profiles could be categorized as follows:

- *Predictable / Foreknown*: The traffic has arrival characteristics that can be predicted and therefore become pre-known, for example a guaranteed bit rate, a constant message size (payload), a one-shot message or several messages associated to one traffic event.
- *Sporadic / Non-constant*: The traffic has no systematic arrival characteristics.

Activity state: Alongside the traffic profile, this traffic property will also play a key role to determine whether it would be beneficial to move a UE to (or keep it) in RRC Connection (as in case of high UE activity), and move away otherwise.

- *Low activity to high activity***:** A high(er) activity is characterized by a large(r) payload size and / or by high(er) frequency of payload arrival.

UE context properties:

Mobility profile: A mobility profile which determines fast crossing of cell borders, results in a large burden for the network in terms of mobility-based signalling overhead if the UEs are in RRC CONNECTED state. Such signalling is both towards the RAN (e.g. due measurement reporting) and towards the core network (e.g. due to changes of the mobility management network entity serving the UE). Therefore, the network should avoid keeping high mobility UEs in RRC CONNECTED, if possible. Similarly, UEs in RRC INACTIVE with a high mobility state will more likely move outside the RAN Notification Area (RNA). This is at least valid under the assumption that the RNA will not be configured too large, in order to avoid extensive signalling to retrieve the UE context, which is required in the resume procedure. Therefore, additional state transitions to RRC CONNECTED will be performed to support the location update procedure (i.e. RNAU). Thus, in case of high RNAU frequency, the use of RRC IDLE mode may be preferred rather than RRC INACTIVE.

- *UE mobility state***:** The mobility state of the device or end terminal can be expressed with the required / available granularity such as stationary, low, medium, high, very/ultra-high.

- *UE velocity***:** For each mobility state given above, a certain speed range could be assumed for the terminal. For instance, the high mobility state could be mapped to a high-speed vehicle, and similarly a very/ultra-high mobility state could be mapped to a high-speed train.

- *UE trajectory***:** The UE trajectory denotes the trajectory or route that the UE will be covering in the upcoming period. This information can be estimated or pre-known according to the road route planner of a vehicle, or pre-known in case of e.g. train routing. Such knowledge could aid the definition of a UE specific network configuration.

Based on the above considerations, the initial framework for RRC state handling is exemplified in Table 3-1, showing qualitatively how (a) the data activity, (b) the UE power availability

properties, (c) the network control-plane load, and (d) UE mobility should influence such handling. Note that the grey shaded entries denote state transitions which are not meaningful, because not feasible or not relevant.

Table 3-1: Summary of the RRC state-handling framework.

		UE moves to / remains in RRC CONNECTED	**UE moves to / remains in RRC INACTIVE**	**UE moves to / remains in RRC IDLE**
UE in RRC CONNECTED	Data activity	High Activity UE Power Available Low/medium CP Network Load	Low Activity UE Power Restriction High Network CP Load	(Not relevant)
	No data activity	Known high activity profile UE Power Available Low/medium CP Network Load Low mobility state	UE Power Restriction High Network Load Low/Medium mobility state	Known very low and sporadic activity High mobility state
UE in RRC INACTIVE	Data activity	Data amount cannot be transferred without state transition	Data amount can be transferred without data transition Infrequent RNAU	(Not relevant)
	No data activity	(Not relevant)	Known activity profile Low/medium CP Network Load	Known very low/ sporadic activity UE Power restrict. High freq. of RNAU
UE in RRC IDLE	Data activity	Always	(Not feasible)	(Not relevant)
	No data activity	(Not relevant)	(Not feasible)	Always

In the future work, we will prove how the framework leads to enhanced RRC performance, as well as, we will explore further enhancements relying on the UE and network context, for instance analyzing how to optimize the RRC state transitions based on the knowledge/estimate of e.g. the UE mobility, traffic profile and network load.

3.1.2 RRC Design for Multiple Network Slices

The concept of network slicing has been introduced in previous sections. One of the major challenges faced when extending network slicing to the fixed and wireless access areas is the vastly different requirements. In this regard, the functional challenges can be summarized as:

- Concurrent support of a large variety of use cases with diverging requirements,
- Support of multi-connectivity to a variety of network slices
- Forward compatible to support new services.

Different RRC parameter configuration sets need to be defined for these services and following alternatives can be considered:

- Triggering conditions of RRC state transition are the same for all network slices;
- For each RRC state, different RRC configuration parameters are chosen, e.g., DRX timers per network slice;
- The combination of above approaches.

The advantage of the second approach using only different RRC configurations is that it will cause minimal specification impacts and less signalling even though the flexibility might be limited.

Therefore, the second approach is adopted in this work. We employ a multi-stage approach to investigate this problem. Starting from per UE traffic analysis in terms of delay (defined as average delay of N packets transmitted within one transmission time window T_p), energy consumption, etc. for a given service belonging to a certain network slice, specific requirements will be identified and then the RRC parameters will be optimized. Here one of the main focuses is the DRX parameters and the objective of optimization is to minimize the energy consumption but at the same time to satisfy certain specific service requirements. As the optimization depends largely on the per-UE traffic profile, the aim in D3.1 is to characterize analytically a generic traffic profile which will be described in the following.

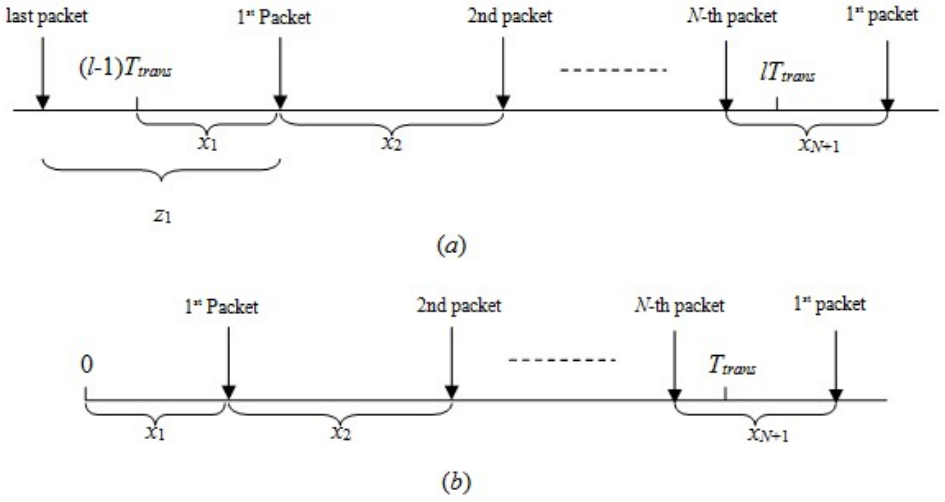


Figure 3-1: Traffic model

Assuming Poisson process for event based traffic with mean arrival rate $\lambda=1/T_{ini}$ as shown in Figure 3-1(a), we consider the general case where N packets are assumed to arrive between time $(l-1)T_p$ and lT_p (T_p is described as T_{trans} in Figure 3-1).The n-th packet arrives at time

$$s_n = (l-1)T_p + \sum_{i=1}^{n} x_i, \tag{1}$$

where x_i is the inter-arrival time between packet (i-1) and i except x_1. According to Annex 6.1, the delay of the n-th packet d_n has no relevance to l so that we can consider a simplified case as shown in Figure 3-1(b).

If N=1, i.e., there is only one packet arrived during time [0, T_p). The joint density for X_1 and S_2 is given in Annex 6.1 as

$$f_{X_1S_2}(x_1,s_2) = \lambda^2 \exp(-\lambda s_2), \quad \text{for } 0 \le x_1 \le s_2. \tag{2}$$

Obviously, the joint density does not contain x_1. Thus, for a fixed s_2, the conditional density of X_1 given $S_2 = s_2$ is uniform over $0 \le x_1 \le s_2$. Considering N=1, it implies that $T_{trans} \le s_2$ so that the conditional density of X_1 is also uniform over $0 \le x_1 \le T_P$. It is easy to see that the delay d_n also follows the uniform distribution over [0, T_P).

For the more general cases, the same behaviour is observed here as

$$f_{S_1 S_{N+1}}(s_1,\ldots,s_N,s_{N+1}) = \lambda^2 \exp(-\lambda s_{N+1}), \quad \text{for } 0 \le s_1 \le \cdots s_N \le s_{N+1}. \tag{3}$$

The joint density does not contain any time of arrival other than s_n, except for the ordering constraint $0 \le s_1 \le \ldots \le s_{N+1}$, and thus this joint density is constant over all choices of times of arrival satisfying the ordering constraint. If any s_n is uniformly distributed, the delay d_n is also uniformly distributed and the pdf and CDF functions are, respectively,

$$p(d_n) = \frac{1}{T_p}, F(d_n) = \frac{1}{T_p} d_n, \text{ for } 0 \le d_n \le T_p. \tag{4}$$

This equation reveals a very important conclusion that the distribution of the delay is solely determined by T_p and bears no relevance to the inter-arrival time. As we know, inter-arrival time is one of the most important parameters to describe a certain traffic type, the above equations mean that the distribution of delay does not depend on mean inter-arrival time but only on T_p, which represents scheduling cycle, where the scheduling decision is valid.

Energy consumption and signalling overhead are essential to the network slice providing mMTC services since the devices are expected to a last long time with battery and the information payload is normally small. However, for other network slices, e.g., slice providing URLLC service, these requirements are not essential. For example, for one of the URLLC services – Safety information in V2X – reliability is more important. One typical scenario is High Density Vehicle Platooning (HDVP), where platoon control could be either centralized or distributed with various degree of control assigned to the leader. Low latency reliable connectivity is an enabler of the platoon application to ensure stable string of vehicles with reduced time headway between them.

HDVP can be formed in advance or dynamically while the vehicles are still moving. For the latter case, low latency and high reliable RRC connection establishment is very important in order not to cause traffic chaos and safety issues. In order to establish an RRC connection, each vehicle needs to send a random access preamble to the base station and there is a chance that two preambles collide so that the vehicles need to re-send them. In this regard, it will cause long latency so that high reliability is needed with extremely low collision probability.

According to [GJK+16], the requirement bandwidth to support N_v vehicles can be expressed as

$$B = \frac{L_{pkt} R_{gen} N_v}{S_{mcs} \eta} \ , \tag{5}$$

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 and η is defined as access efficiency, describing the level of access coordination. η ranges between 0 and 1 and η=1 means perfectly coordinated access and η=0 means uncoordinated access.

If we assume each vehicle needs to transmit a preamble to the base station in full coordination, all preambles can be put in a queue as long as $N_v \le N_s$, where N_s is the number of time slots for all vehicles, and therefore no collision happens, i.e., reliability requirement can always be met. The average latency can be given as below

$$T = \sum_{N_v} \left[1 - \left(\frac{N_s - 1}{N_s} \right)^{n-1} \right] \Pr(N_v = n) \frac{n L_{pkt} R_{gen}}{S_{mcs} B} . \tag{6}$$

For network providing MTC service, further analysis will be done for other parameters such as energy consumption, signalling overhead, etc. in future works and how to optimize specific RRC parameter sets for RRC state transition in different network slices will be studied. For network slice providing ultra-reliable service, full coordination is not always possible in reality. In future

works, we will also take none or low coordination cases into consideration and analyse the performance of different schemes taking WP2 use cases into consideration.

3.1.3 Optimized DRX handling for bandwidth adaptation in 5G NR

Based on the NR DRX framework described in Section 2.3, in this section we present several methods for BWP inactivity timer configuration in the context of BWP switching. Specifically, Method 1 aims at having large BWP inactivity timer value so as to reduce the number of BWP switching and the associated signalling overhead. Method 2 is designed to have small BWP inactivity timer so as to minimize the transmit/receive time of large bandwidth. Moreover, for Method 2, default BWP based and non-default BWP based retransmission scheduling schemes are further detailed. In case of non-default BWP based retransmission, one-step and two-step BWP switching approaches are also 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 optimize the overall UE power consumption in the context of bandwidth adaptation.

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, the *bwp-inacitivityTimer* 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.

BWP-specific DRX and inactivity timer

In case of different numerologies applied in different BWPs, the DRX timers including *HARQ RTT* and retransmission timer can be configured in BWP specific manner, i.e., each BWP shall have its respective timer configuration. In this case, *bwp-inactivityTimer* shall be configured in BWP specific as well.

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, UE can switch to the default BWP earlier than the retransmission timer expires. If this happens, the following scheduling options can be used for UE to receive the retransmission packet.

Option 1: retransmission in default BWP

If code-block-group (CBG)) based retransmission is used, it is possible to retransmit only one CBG with a smaller packet size which can be transmitted in the default BWP. In this option, the retransmission happens in the default BWP within the retransmission time window, i.e., before retransmission timer expires. This would achieve better energy efficiency by minimizing the unnecessary usage of large bandwidth.

Option 2: retransmission in non-default BWP

In case of non-CBG based retransmission, i.e., transport block (TB) based retransmission, the retransmission packet has the same size as the initial transmission. This could require retransmitting the packet in the non-default BWP where the initial transmission happened. In this case, before gNB schedules the retransmission packet in the non-default BWP, gNB shall first request UE to switch to the previous non-default BWP. The following two alternatives can be used.

Alternative 2.1: One-step approach

In this approach, gNB uses the DCI transmitted in the default BWP to schedule the retransmission in non-default BWP. Since UE needs some time to switch the active BWP, so cross-slot scheduling needs to be used. Alternatively, one or several OFDM symbols interval between the PDSCH start symbol and the last symbol of PDCCH should be reserved for UE BWP switching. This assumes the supported range of such time interval should be signalled by UE to the network as part of UE capability.

Alternative 2.2: Two-step approach

In this approach, gNB shall first send to the UE a DCI without PDSCH scheduling to only request for BWP switching. After UE acknowledges the request, gNB shall schedule the retransmission in the active large BWP. As such, this approach involves two signalling steps. This would require a large retransmission time window.

Method 3: Efficient BWP inactivity timer configuration

As mentioned above, BWP inactivity timer can have different settings with respect to HARQ retransmission time window, it is up to the network to choose which options to be used. Different options 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, it is envisioned that there is small possibility for retransmission, it may be good to have 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 signalling for BWP switching. This also depends on the energy consumption comparisons between BWP switching signalling and wideband BWP operation. If the former is smaller, smaller BWP inactivity timer should be configured; otherwise, large BWP inactivity timer can be used. More detailed quantitative evaluation for different methods can be studied in future work.

3.2 Multi-Service Resource Allocation Optimizations

This section summarizes the ongoing research and current findings related to multi-service resource allocation optimizations, and hence has its main focus on MAC-layer scheduling functionalities. The following techniques are in focus:

1. Pre-emptive scheduling methods for efficient mux of eMBB and URLLC.
2. Resource allocation methods for cases where network slicing is applied.
3. MEC-aware and prediction-based resource allocations methods
4. Centralized scheduling methods and efficient allocation of channel quality indicator (CQI) reports is outlined.
5. Prediction and CQI scheduling techniques.

Those techniques are complementary and relevant for different network architectures and use cases. Techniques 1 and 2 address multiplexing of diverse services using two different approaches (without and without network slicing), while 3 and 4 are optimizations specifically tailored for MEC and C-RAN architectures.

3.2.1 Pre-emptive scheduling for MUX of eMBB and URLLC

In this section, we summarize our studies of efficient multiplexing of eMBB and URLLC traffic types on the downlink shared channel as e.g. is relevant for ONE5G use case no. 2 [D21]. This is a challenging problem, given the diverse QoS requirements for eMBB and URLLC. Focus is on MAC-layer scheduling functionality and optimization, where some users have DRB(s) (Data Radio Bearer) with best effort eMBB, while other users have DRB(s) with the strictest URLLC requirement of 1 ms latency and 99.999% reliability. In line with earlier studies, we consider the case where different TTI sizes are used for eMBB and URLLC to achieve the best possible

performance. Use of different TTI sizes for the NR has also been studied earlier in [LBL+16][PNS+16][PPS+16][PSP+17]. The technique studied for multiplexing of eMBB and URLLC is named pre-emptive scheduling (aka punctured scheduling). This method is adopted also by 3GPP for the NR. This scheme has similarities to pre-emptive scheduling principles as studied extensively for computer networks to accommodate real-time services. The basic principle is as follows: eMBB users are scheduled on all available radio resources. When latency-critical traffic suddenly appears at the gNB, the data is immediately transmitted with a short TTI that fully or partly overwrites a fraction of an ongoing eMBB transmission that uses a longer TTI size. By doing this, there is no need for reserving radio resources for randomly arriving URLLC traffic. The cost of pre-emptive scheduling is on the eMBB performance as being partly overwritten. To minimize the eMBB performance loss from pre-emptive scheduling of URLLC payloads, several recovery mechanisms are considered. Those include:

1. Pre-emptive indication (PI)[1], where the gNB makes the victim eMBB aware of which resources of its transmission have been overwritten by a pre-emptive scheduling transmission.
2. Smart HARQ retransmission schemes where e.g. only the damaged part of pre-empted eMBB transmissions are retransmitted in case of failures.

In our first study (published in [PPS+17]), the performance of pre-emptive scheduling is studied for cases with/without PI, considering different RRM methods for scheduling policies and dynamic link adaptation (LA). A service-dependent LA scheme is applied, where the initial first transmission BLER target for eMBB and URLLC is set to 10% and 1%, respectively. Standard Proportional Fair (PF) scheduling is used for eMBB resource allocation, assuming a 14-symbol TTI size. URLLC is scheduled with a short TTI size of one 2-symbol mini-slot. By doing this, multiplexing of different TTI sizes is achieved, while still using a single unique PHY-layer numerology of 15 kHz sub-carrier spacing, for the assumed macro-cellular environment. Three different scheduling policies are studied for URLLC transmission, including so-called eMBB-aware schemes, where either radio resources used by eMBB users served with low or high modulation and coding scheme are prioritized for pre-emption. A simple Boolean HARQ ACK/NACK feedback is assumed, where the gNB upon reception of NACK retransmits the full transport block. The performance is studied by means of advanced macro-cellular system-level simulations, assuming full buffer traffic for eMBB and Poisson arrival of 50 bytes URLLC payloads. The main findings are (for details, see [PPS+17]):

- The obtained system-level performance results highlight the benefits of pre-emptive scheduling, confirming our hypothesis that pre-emptive scheduling is attractive and worth applying for mux of eMBB and URLLC. It provides clear benefits for the URLLC defined KQI service integrity counter (as defined in [D21] since the latency is minimized.
- Pre-emptive scheduling is generally found attractive, as it does not require any pre-reservation of radio resources for transmission of the randomly arriving URLLC payloads.
- Using the PI signaling offers attractive gains.
- Service-specific and pre-emptive-aware dynamic LA, as well as eMBB-aware scheduling decisions are recommended. Particularly, it is advised to prioritize overwriting resource elements for eMBB users transmitted with low modulation and coding schemes (for cases where no free resources are available for immediate scheduling of URLLC).

Following our study in [PPS+17], additional enhancements are developed and studied, including more detailed assessment of eMBB E2E performance because of experiencing pre-emptive scheduling. The latter is achieved by using a more realistic finite buffer file-download traffic for eMBB, including the Cubic TCP model where e.g. the TCP slow start mechanism is modelled. This allows to study how pre-emptive scheduling influence on the TCP performance (e.g. does it

[1] Notice that PI is now also referred to as "Interrupted Transmission Indication" in the latest version of 3GPP TS 38.214.

cause undesired TCP slow starts that harms the end-user performance?). The enhanced recovery and HARQ schemes listed in Table 3-2 are studied. HARQ scheme #1 is identical to what we used in [PPS+17], while scheme #2 include an enhancement where upon reception of NACK for an eMBB transmission that has been subject to pre-emption, only the damaged part of the original transmission is first retransmitted (for the first reTx). Hence, we utilize the knowledge at the gNB of which resource elements of the pre-empted eMBB transmission have been overwritten. For HARQ schemes #3 and #4, we propose using multi-bit HARQ feedback to have one ACK/NACK per code block (CB) that appears in the eMBB transport blocks. We consider both the cases with fully interleaved (random) CB layout, and also the so-called frequency-first CB layout. For HARQ scheme #4, only the pre-empted part of the CBs where a NACK is received at the gNB is retransmitted (for the first reTx). PI signalling is enabled for all cases. The findings from these studies can be summarized as follows (details in [PPS18]):

- At the medium fractile (50%-ile), the degradation of TCP layer throughput, and increased round trip time (RTT), is only on the order of 20% when carrying an additional latency critical traffic of 2 Mbps/cell, where every payload of 50 bytes is timely delivered within 1 ms at an ultra-reliability level of 99.999%.
- This in line with theory: with no latency constraints, the effective capacity equals the Shannon capacity, while it decreases asymptotically as stricter latency constraints are enforced.
- Using pre-emptive scheduling is therefore a feasible solution, even when considering the influence on the eMBB E2E performance, and particularly the impact on TCP. Relating to the KQI definitions in [D21] for file transfer, this maps to the service integrity category, resulting in a longer file download due to the pre-emption.
- The configuration with the fully interleaved CB layout generally offers the best performance, while the frequency-first option only is marginally better if using HARQ scheme #3 (as per Table 3-2)

For the future WP3 deliverable, we will further study enhanced schemes for multiplexing of eMBB and URLLC by utilization of advanced MU-MIMO scheduling schemes for cases where the gNB is equipped with at least eight antenna ports. Our hypothesis is that this opens for derivation of new spatial multiplexing schemes of eMBB and URLLC, rather than having to fully pre-empt eMBB transmissions when urgent URLLC payloads needs to be scheduled. First findings from such studies are reported in [EP18].

Table 3-2: Studied recovery and HARQ schemes.

HARQ scheme	Description	UE feedback
#1	Baseline: Upon reception of Nack, the full TB is retransmitted.	Single-bit ACK/NACK
#2	Partial retransmission: if the TB was subject to preemption, only the damaged part of the TB is retransmitted. Only if a second NACK is received for the same HARQ process is the full TB retransmitted again.	Single-bit ACK/NACK
#3	Only the CBs with NACK feedback are retransmitted. This is valid independently on whether the TB was subject to pre-emptive scheduling, or not.	CB-based ACK/NACK (multi-bit feedback)
#4	If the first TB transmission was subject to preemption, only the damaged parts of the CBs having received a corresponding NACK for is retransmitted. If full TB is not correctly decoded after the second HARQ transmission, the full TB is retransmitted again. If the first TB transmission was not subject to pre-emption, HARQ scheme #3 is applied.	CB-based ACK/NACK (multi-bit feedback)

3.2.2 Efficient resource allocation for network slices

The contribution in this section focuses on Network Slicing, as described in section 2.2. The main principle in network slicing is that related services can be assigned to a common network slice, for which the operator can decide the amount of allocated resources as well as scale up/down the resources when services join/leave as well as reject new services in cases where QoS may be compromised if additional services are allowed to join the slice. Another important aspect of network slicing is the isolation between slices, meaning that shortage of resources in one slice must not affect the performance of other slices, unless the network operator chooses to reallocate resources differently among slices. Since network slicing aims to ensure coexistence of various service types, we are mainly considering the use cases 1 (assisted, cooperative and tele-operated driving) and 2 (time-critical factory processes and logistics optimization) from [D21] where all considered services categories (URLLC, eMBB, and mMTC) are present, since the aim of network slicing is to ensure the co-existence of various different services.

An important component of the network slicing framework is therefore the ability of the network operator to foresee the achievable performance, given a certain amount of resources to serve a specific set of services in a network slice. Network slices are typically defined as end-to-end, meaning that all necessary resources between the two end-points of the network slice needs to be accounted for and provisioned (at least statistically) according to the service requirements.

In the present work item, we have initially focused on establishing the model that will be used to predict the amount of resources needed to satisfy the performance requirements of different types of traffic. The proposed model is based on queueing model theory, and considers each network slice as a server that is able to serve the arriving traffic. In the simplest case of network slicing, where dedicated resources are allocated to different network slices and it is not possible to share resources, the different slices can obviously be modelled separately. An example of this is the in-band deployment of NB-IoT in dedicated resources inside the LTE frequency band. However, such an approach leads to a poor utilization of resources. Ideally, in 5G NR, resources should be shared between slices to allow for multiplexing and efficient utilization of resources, while ensuring that experienced performance of users lives up to the agreed service level agreements. Modelling this sharing of resources is not trivial, since the resources available to serve traffic in one slice, depends on the arriving traffic in other slices. Having not been able to find a suited model in the state-of-the-art literature, we propose to model the system with shared resources between slices as follows.

For each network slice, we consider the arriving traffic as the arrival rate λ and the available resources (Physical Resource Blocks, PRBs per time unit) as μ. Further, the latency limit is quantified as a maximum tolerated queue length.

Figure 3-2: Queue model with maximum queue length indicated.

Based on this model, illustrated in Figure 3-2, the reliability for latency limit l is calculated as:

$$P_s(l,\lambda,\mu) = 1 - \sum_{i=0}^{l} \pi_i\,(\lambda,\mu),$$

where $\pi_i(\lambda,\mu)$ is the steady-state probability of state i of the queue model defined by arrival rate λ and service rate μ.

Assuming that slices are served according to a prioritized order, we calculate the needed resources for slice j, expressed as service rate μ_j, by the minimization:

$$\min_{\mu_j}\left[\log_{10}(1-P_s(l_j,\lambda_j,\mu_j))-\log_{10}(1-R_j)\right]^2$$

where λ_j is the arrival rate of traffic for slice j, and R_j and l_j are the reliability and latency requirements, respectively.

In order to support the reuse of unused resources from higher priority slices, the actual allocated resources, in terms of service rate of slice j, for j>1 is defined as:

$$\mu_j' = \max\left(0, \mu_j - \sum_{k=1}^{j-1}\mu_k - \lambda_k\right)$$

For evaluation of the proposed model, we have considered a simplified system model, where only the data plane resources in the air interface are considered for different network slices. For this preliminary study, we assume Poisson arrivals and an LTE-like resource grid with fixed 15 kHz subcarrier spacing, 0.5 ms slots, i.e. no mini-slots. Two example results are shown in Figure 3-3 and Figure 3-4. The figure shows the more realistic case where network slices' unused resources in a timeslot can be used by lower priority slices, thereby achieving high efficiency. This is also evident from the figure, since on the right side, all services achieve at least as good or better performance than on the left side.

Besides the example results in Figure 3-3 and Figure 3-4, tests are being conducted with simple traffic models that use combinations of deterministic or exponential inter-arrival times and payload sizes.

In the future work we are planning to integrate the proposed prediction model in an admission control algorithm, which we will evaluate in a simulated dynamic environment with joining and leaving users. We are currently considering how to take into account also control plane resources. Also, more realistic traffic models will be considered as they are being decided upon/developed in WP2.

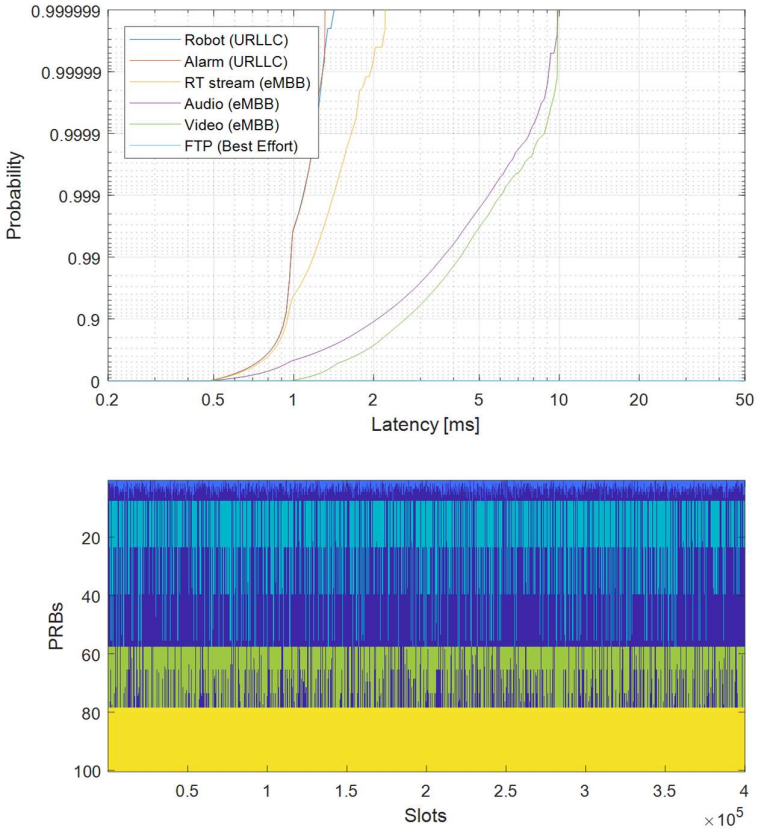


Figure 3-3: Dedicated resources per slice.

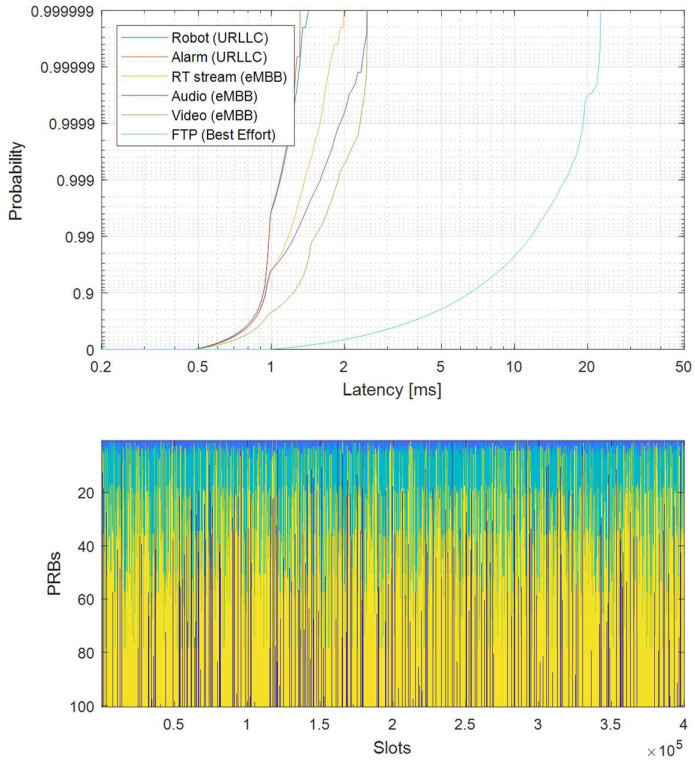


Figure 3-4: Prioritized multiplexing of slice resources.

3.2.3 Time-variant optimal slicing negotiations

This work is motivated by the fact that vertical entities (e.g., utility provider) can benefit from automated negotiation mechanisms in order to get lower prices for the needed resources of a slice. In general, negotiation is the primary form of interaction of two or more parties for the formulation of an agreement [LAH04], [SKD+05],[JPL+01]. The work focuses on mechanisms for the automated negotiation of price offers for providing certain quality levels of services for "megacities" and "under-served areas" scenarios at certain price levels by taking into account environment heterogeneity, variable user aspects (e.g. density, distribution etc.), variable service/traffic demand (e.g. accommodating eMBB, URLLC, mMTC) and network aspects (e.g. cell layout, bands etc.).

The optimization process results in a selection of objective function values (OFV) comprising utility and costs for specific quality levels of the requested traffic sources. The calculation of these allocations optimises an objective function, which is associated with the quality levels at which each service will be provided. A formulation of the problem shall take into account the fact that *given*:

- Set of traffic sources T
- Set of QoS Q(t) where t in T
- Set of costs C(q) where q in Q(t)
- Set of utility U(q) where q in Q(t)

Find negotiated level of $OFV_{vertical}$ and $OFV_{operator}$ so as to serve the required traffic sources with a certain quality at a given cost. The vertical entity shall negotiate for costs and utility of resources associated to quality levels. As mentioned also in [LAH04] the messages exchanged between two negotiators include a set of predetermined issues with their values. In our case, a proposal can be a set of value pairs $OFV_{vertical}(U(q),C(q))$ and $OFV_{operator}(U(q),C(q))$. Final agreement should be Pareto optimal. A Pareto optimal solution that dominates the final agreement should be preferable to both negotiators. Also, list elements are arranged from the most preferable to the least preferable. An updated version of the algorithm in [LAH04] for finding Pareto optimal solutions is available below:

1. Select element from a list of $OFV_{vertical}$ of the vertical entity (e.g. element ev_i);

2. Find the element from a list of $OFV_{operator}$ of the operator (e.g. element eo_j);

3. Get the set PESet1 of previous elements of ev_i till ev_{i-1} ;

4. Get the set PESet2 of previous element of eo_j till eo_{j-1} ;

5. Check whether the intersection of PESet1 and PESet2 is empty. If the intersection is empty, insert ev_i to the end of output list 1 (of vertical entity). Otherwise ev_i (also eo_j) is dominated by other elements, so it will not be added;

6. Copy output list 1 to output list 2 (of operator), reverse elements in output list 2.

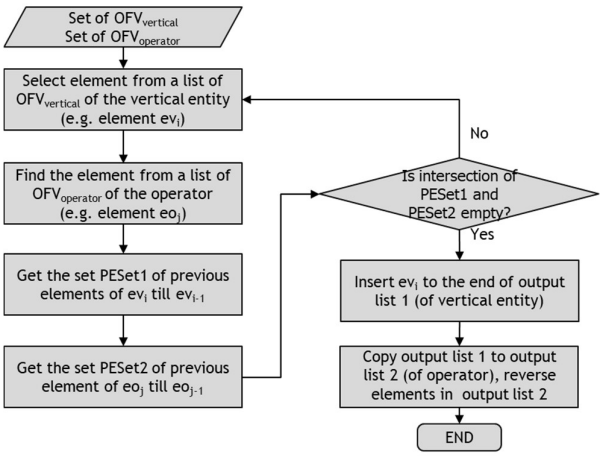


Figure 3-5: Algorithm for finding Pareto optimal solutions.

The aforementioned solution will lead to improved costs and quality for vertical entities, by taking also into account what operators can offer. Preferred (and Pareto optimal) values of vertical entity and operator may differ for each case hence a final selection from the set of preferred values is essential. The following algorithm is analyzed below:

1. The vertical entity proposes a first alternative negotiated value from the list of OFV;

2. The counter that keeps a record of how many times a negotiated value is considered is reduced by 1;

3. The operator receives the proposal;

4. Check whether the pointer that the operator is willing to give in this round matches the pointer of the proposed negotiated value by the vertical entity. If it matches, then an agreement is reached and the algorithm finalizes. Otherwise, it continues.

5. Then the operator proposes a first alternative negotiated value from the list of OFV;

6. The counter that keeps a record of how many times a negotiated value is considered is reduced by 1;

7. The vertical entity receives the proposal;

8. Check whether the pointer that the vertical entity is willing to give in this round matches the pointer of the proposed negotiated value by the operator. If it matches, then an agreement is reached and the algorithm finalizes. Otherwise, it continues with steps 1-4.

The aforementioned work will continue towards IR3.2 and D3.2 in order to provide concrete evaluation results. Such a solution would be beneficial for various use cases either in megacities or underserved areas where vertical entities (e.g. utility operators) can negotiate prices for certain quality levels that can be provided by network slices.

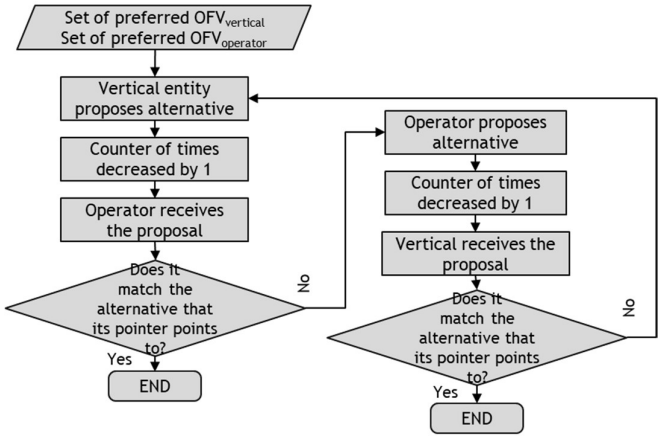


Figure 3-6: Algorithm for final selection from a set of preferred Pareto optimal solutions.

3.2.4 MEC aware resource allocation principles

The increasing demand of various wireless applications running at user devices has recently pushed the boundaries of the users' QoE in an extensive manner. Services and applications for daily purposes, as well as high resolution sensors, virtual reality and complex processes impose stringent requirements on the devices, from an experienced latency, energy consumption and power efficiency standpoint. Additionally, one of the key challenges of these applications is their high computational (processing) requirements. However, the local computation capabilities of mobile devices are still limited, as compared to the emerging application demands. To overcome this limitation, MEC is proposed as a promising solution towards relieving wireless devices from a heavy computational (and, thus, power-hungry) burden via providing processing power at the edge of the network.

Throughout the planned work, focusing on a MEC-enabled Heterogeneous Network (HetNet) which may be characterized by significant inter-tier resource disparities, we aim to investigate the joint allocation of radio and computational resources to devices running multiple services of dissimilar requirements. To accomplish that, we identify three key factors influencing user QoE; (1) allocation of radio bandwidth, (2) allocation of MEC (processing) resources, and (3) selection of bias values for cell association. The bias values (e.g. used for cell range extension in LTE Rel.12), are artificial values utilized to balance the user load in a HetNet. To the best of our knowledge, current technical literature only sheds light on the problem of optimally allocating radio and computational resources to the users, when the latter are *already* associated to a cell. For instance, in [LAK17], the problem of radio and computational resource allocation over connected users was investigated under Time Division Multiple Access (TDMA) and Frequency Division Multiple Access (FDMA) schemes and compared to a baseline round-robin scheme. Moreover, [MZL17] studied the problem of joint power and computation allocation under an optimization framework, where the task completion time was minimized subject to energy consumption constraints. As multi-service QoS provisioning in MEC-enabled deployments has not yet been studied to the best of our knowledge, in this work we aim to accommodate diverse QoS classes focusing on the experienced latency. An optimization framework is currently under construction that aims at computing the optimum radio and MEC allocation portions together with the optimal bias values for cell selection.

3.2.5 Centralized multi-cell scheduling

Centralized RAN, also referred to as C-RAN, is one of the key components that will be part of 5G Standard from its first release (NR Release 15) due to the innovative element in the architecture which allows CU/DU split. As was highlighted in the previous sections, C-RAN deployments have a huge impact on the cellular network architecture especially in terms of cost

and information exchange between the nodes, since hardware centralization allows reducing the operational and maintenance costs as well as the amount of signalling information exchange between cells to coordinate transmissions, which becomes unmanageable in ultra-dense environments.

In this work, we focus on C-RAN implementations with Split points at intra-MAC level (or below), such that the MAC scheduler entity is centralized. The basic idea behind the analysed C-RAN deployment is that of a "super-cell" being managed by a CU, performing all radio tasks above the Split point corresponding to the different cells, while leaving the remaining lower-layer RAN functions at the Remote Radio Heads (RRHs). Figure 3-7 illustrates the C-RAN concept, where CU comprises the baseband processing unit for a number of cells and the RRHs at the sites (equipped with one or more antennas) perform the remaining RAN functions below the Split point.

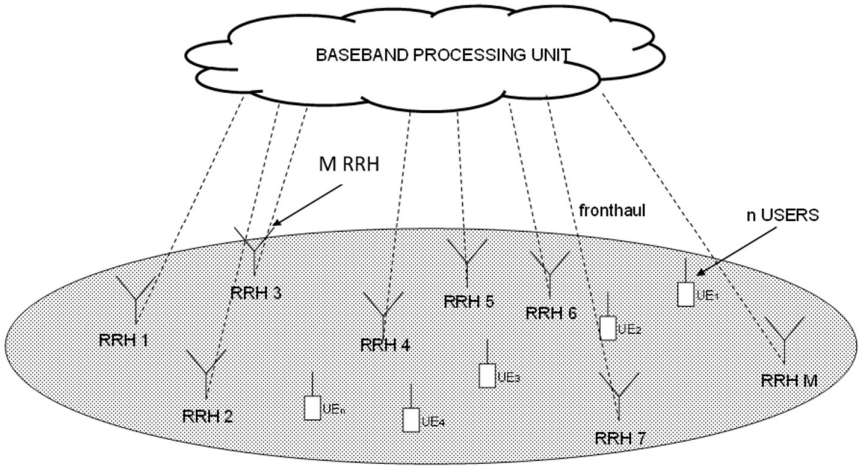


Figure 3-7: C-RAN Deployment.

As can be noticed, ultra-dense environments comprising hundreds of cells and thousands of users can pose a challenging problem when trying to schedule resources efficiently, e.g. because of the presence of numerous interference sources. In this section, a centralized multi-cell scheduler is proposed that solves the problem of how to schedule resources efficiently in a complex centralized radio access network, with the intention to maximize the overall scenario capacity and minimize the interferences.

By taking advantage of the measurements that the users report to the gNB, such as the Channel Quality Indicators (CQIs), the centralized multi-cell scheduler will allocate users to Resource Blocks (RBs) in an efficient way through selecting the users and the bandwidth portions where channel conditions are the best. Nevertheless, the centralized multi-cell scheduler will not penalize the users at cell-edge, but instead still assign resources to them even when their channel conditions are not so good for the sake of fairness amongst the users, as proportional fair schedulers used in 4G systems.

The centralized multi-cell scheduler will use a table populated with Proportional Fair (PF) metrics to schedule users at the available sub-bands and RRHs. All the RRHs have the same cell identifiers as they are connected to the same CU hence avoiding handovers. Having the same cell ID allows skipping the initial random access process every time the scheduler selects another RRH, since the UE is always logically connected to the super-cell even if different RRHs can transmit to the UE at different time instants. Moreover, this allows reusing radio resources when there is enough RF isolation between the RRHs involved, and allows further application of advanced coordination techniques such as Coordinated Multipoint (CoMP) and Non-Orthogonal

Multiple Access (NOMA). Both CoMP and NOMA techniques will be taken into account by the C-RAN scheduler.

Figure 3-8 shows the 3D table that the scheduler will use to allocate UEs to the different sub-bands and RRHs. The selected sub-band size is arbitrarily chosen to be equal to four RBs, so that a 20 MHz system bandwidth will comprise 25 sub-bands.

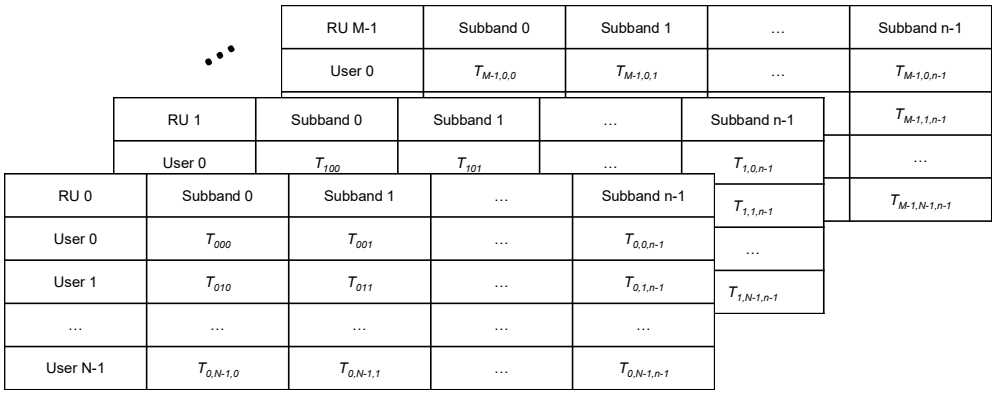


Figure 3-8: 3D Scheduling Table.

Users can make use of the whole bandwidth as granted by the scheduler. However, it is possible to put additional restrictions in the form of a maximum number of scheduled sub-bands per UE. This can help for latency-sensitive services not to spend a long time waiting to be served in the MAC queue, as would happen otherwise in systems without similar restrictions.

This work proposes a suitable time-frequency algorithm that is applicable to OFDMA networks for the Megacities scenario. In OFDMA systems, several resource allocation problems with given constraints are NP-hard, which makes them intractable when the number of sectors, users and/or subcarriers grow. The proposed centralized multi-cell scheduler is a sub-optimal alternative with low complexity, equal to $O(N \cdot M \cdot L)$, where N is the number of users, M is the number of sectors, and L is the number of subbands. Traffic handled by the scheduler is assumed to be mostly eMBB, albeit with delay restrictions as per the maximum number of scheduled sub-bands per UE. Subcarrier spacing and transmission time interval are assumed equal to those in LTE, i.e. 15 kHz and 1 *ms*, respectively. However, the scheduler can also serve latency-sensitive traffic by avoiding excessive queueing delays at the MAC layer.

UL and DL procedures are decoupled in the scheduler, i.e. UL scheduling decisions are independent to those at the DL. The optional presence of DFT-spread -OFDM (DFT-s-OFDM) modulation in the UL imposes an adjacency requirement for the scheduled frequency resources, so that the scheduler will search for contiguous sub-bands maximizing the overall capacity. Nevertheless, for DL there will be no adjacency restrictions in the scheduled resources, thus making it possible to allocate UEs to any suitable sub-band.

Besides the 3D scheduling table, C-RAN scheduler will have information available related to performed allocations, such as:

- a set of users containing the users that have been assigned a total K of subbands in RRH *i*, where *K* can be equal or less to *L*, i.e 25 subbands

- a set of sectors containing the RRHs have not been yet scheduled and,

- a set of subbands containing the available subbands within the considered subframe.

Therefore, C-RAN scheduler will work by selecting randomly a subband, *k*, and a sector, *RRH i*, among the set of subbands and sectors not yet assigned to perform the allocation. For the selected subband and sector, DL C-RAN scheduler will find the user, *jo*, with the highest metric in 3D

scheduling table and among the set of available users as a candidate to be scheduled in subband *k* and *RRH i*.

However, centralized multi-cell scheduler aims to boost overall capacity by scheduling the users in subbands and sectors where the channel conditions are the best. So that, C-RAN will study whether it is possible to select another RRH i' and subband *l* where the above candidate, *jo*, has higher metric than in the selected ones by following the next steps.

1. Analyse whether there exists other *RRH i'* for which the user has higher metric among the set of RRHs not yet scheduled and that in turn do not have assigned subband *k*. In that case, the proposed scheduler will swap *RRH i* for *i'* to continue the study.

2. For the selected RRH, check whether there exists another subband *l* different than *k* for which user *jo* has a higher maximum value of the metric, T_{ijk}, among the set of available subbands.

    a. In affirmative case, and to elucidate whether user j_0 might be a better candidate for subband *l* than for subband *k*, analyse secondary maxima as follows:
        i. Find users, *j1* and *j2*, with second highest metric in subband *k* and *l,* respectively. If $T_{ij_1k}+T_{ij_0l}>T_{ij_0k}+T_{ij_2l}$ then user *j0* is a better candidate for subband *l* and *j1* for subband *k*. Otherwise, user *j0* is a better candidate for subband *k* and *j2* for subband *l*.

    b. In the other case, user *jo* is the best candidate to be scheduled in subband *k* and sector *RRH i.*

After studying the above conditions, the algorithm will check whether the assigned subbands can be used by other RRHs by means of applying CoMP, NOMA or RF isolation techniques and repeat the procedure until there is no more RRHs or subbands available.

UL resource allocation follows basically the same principle as DL scheduling explained above except that UL procedure impose the adjacency requirement so that for UL scheduling the algorithm will have to base the study in analyzing a set of adjacent subbands instead of a single one, as done before. Hence, UL C-RAN scheduler will find the user *jo* with the highest metric in a set of *n* adjacent subbands, where *n* is an user-defined parameter, and carry out the above checking considering a set of *n* adjacent subbands.

In next deliverable, the performance of the proposed algorithm will be studied by means of analyzing the obtained results in the System-Level Simulator (SLS) in terms of throughput improvement under different scenarios (Manhattan and Canonical). It is worth mentioning that, in collaboration with WP2, the work in this section is proposed for implementation in WP2 system level simulations.

3.2.6 Prediction techniques for improved routing performance

Optimally routing traffic flows, i.e., selecting the links that will be used to forward data from a source to destination nodes, is a challenging problem. This is especially the case for the dense network topologies considered in megacities use cases such as outdoor hotspots and live events discussed in [D21]. These use cases may involve multi-hop links using relays, e.g. leveraging on the D2D capabilities of UEs, as well as cooperative transmissions with wireless backhaul/fronthaul. In principle, the nodes involved have the ability to establish links with each other and the goal is to identify the optimal information flows among them to achieve a certain objective (e.g., the timely and uninterrupted delivery of a streaming service to a destination node). In this study, we consider the problem of dynamic/adaptive traffic routing in a C-RAN setting, where a CU takes optimal routing decisions based on available information about the network state, characterized by 1) the buffer state of each node as well as 2) the connectivity of nodes (in terms of the probability of successful transmission). The nodes in consideration could be RRHs

and UEs as in a standard C-RAN setting, however, the model is general enough to also cover cases involving other/additional link types such as D2D, where having a CU taking routing decisions for these links as well is expected to be beneficial, in principle.

Clearly, since the state of buffers and channels/links in the system are time-varying, *optimal* routing should be adaptive to their values. In this setting, previously proposed routing schemes, rely only on current and one-step-ahead prediction of the network (channel) state [TA92], [KW18]. However, when the channel/link states can be predicted (in a stochastic sense) far into the future, one expects that exploiting this information can result in improved routing decisions. This idea is developed here, namely, we aim at obtaining the optimal routing decisions by utilizing not only current channel state information but also its prediction, potentially far into the future.

The proposed adaptive routing algorithm evolves around the convex optimization of a utility function, under consideration of the system's state evolution dynamics and the predicted system (channel) behaviour in the future. We consider a general network topology involving n (wireless) nodes, each with its own data buffer containing a certain amount of information to be transmitted at any given time instant. The buffer state of all nodes in the system at time *t* is represented by the queue-vector $q_t \in \mathbb{N}_+^n$ and the (discrete time) system state evolution can be written as [Mey07]

$$q_{t+1} = q_t + B_t u_t + a_t,$$

where $a_t \in \mathbb{N}_+^n$ is the data arrival vector, which follows a generalized Bernoulli process, $u_t \in \{0,1\}^m$ is the routing decision vector at time *t* whose value is decided by the CU and dictates forwarding of packets among the nodes in the system, and the discrete-valued matrix $B_t \in \mathbb{Z}^{n\times \mathrm{m}}$ ($\mathbb{Z}$ denotes integers) represents the network topology, i.e. which nodes can exchange how much information at a given time. In detail, each column vector of B_t is a communication link, which when activated, subtracts information in some queues and adds information to others. B_t also encodes which nodes are destination nodes, i.e. where information leaves the system. In addition, at any given time, not all available communication channels may be allowed to be active (e.g., due to physical limitations). This restriction can in principle be represented by the linear constraint $Cu_t \leq 1$, for an appropriate matrix C.

For wireless networks, sequence B_t is not constant but changes with time due to effects like channel fading or temporary interference. In our novel control approach, we introduce a Markov chain describing this behaviour by changing transmission success probabilities for each link in possibly each time step. We decode these success probabilities in weight matrices $\{M^i\}$ and let the Markov chain σ_t (with initial state σ_0 and transition matrix *P*) evolve on its index set $I\{M^i\}$, so that $\sigma_t = \mathbf{M}\big(I\{M^i\}, P, \sigma_0\big)$. We further use Bernoulli Trials $\mathbf{B}(\bullet)$ to let a single communication succeed or fail, depending on the probabilities in M^{σ_t}. Hence $B_t = B \cdot \mathbf{B}(M^{\sigma_t})$, where B is the constant routing matrix of the system topology.

We implement a *model-predictive* framework for traffic routing by considering a utility function that especially contains predicted channel states. This utility function is of quadratic type, using given diagonal weight matrices Q and R, and contains expectations of future system quantities:

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

Without further consideration the control has to be subject to $Cu_{i-1} \leq 1$ and $q_i \geq 0$ for all $i = (t+1) \ldots (t+H)$, where $H > 1$ represents the number of steps ahead for which the channel state is predicted. Our control policy now consists of solving the above problem to obtain the optimal routing trajectory $u_t^*, \ldots, u_{\{t+H-1\}}^*$ and applying u_t^* to the system at time t. The whole process is repeated at time $t+1$ and so on. Note that the use of a decision-weight matrix R to penalize control decisions is unique to our algorithm and aims at reducing the amount of required control decision, thus saving energy.

Consideration of a future horizon $H > 1$ in the objective function shows improvements in performance in comparison to choosing $H = 1$ which equals the conventional, myopic control approaches. A demonstration of this can be seen in the preliminary simulation results of Figure 3-9 where the right panel shows the time evolution of the buffer state over time in an exemplary network topology shown in the left panel, obtained under the conventional max-weight routing policy (orange curve) [TA92] and the proposed routing policy (blue curve). The considered network is derived from a practical scenario where 2 users have to compute shared information synchronously. Due to bad connectivity, communication link u_1 is only rarely available, which our new control policy considers in its prediction of the future. Note that the less information is stored in buffer 1, the more information flows successfully through the network. The curves show that our new policy outperforms the conventional one in times of over-the-average arrival rate bursts.

Furthermore, our developed *model-predictive* routing is also able to increase the stability region of networks with a more complex structure, illustrated in Figure 3-10. In this scenario, user 1 (corresponding to q_1) has to process data but can take huge advantage of parallel, synchronized data processing with the help of user 2 (q_2). Naturally, the mere transmission of the data to user 2 presents a certain additional cost, since in this time the data is not processed. These costs prevent standard algorithms from fully taking advantage of the parallel processing capabilities of the entire system. In contrast, our new control policy is able to fully use this parallel processing potential, since it realizes that these additional costs are outweighed by the boost in overall performance, enabling greater data input data rates a_1. In general, this implies that we can potentially facilitate a greater overall transmission rate in certain scenarios like communication in between Cyber-Physical-Networks.

A detailed description of this work appears in [SW18b]. Future work will consider the stability characterization of the proposed approach as well as determination and reduction of computational requirements. The performance of the algorithm will be evaluated also in terms of other metrics of interest such as latency and long-term throughput where gains are expected to be observed even though not explicitly optimized by the considered formulation.

It is noted that the proposed routing mechanism is aligned to the current status of C-RAN architecture considered in 3GPP and the decided higher layer split as described in Sec. 2.2. In particular, the proposed mechanism is naturally suited to a PDCP level implementation, where routing decisions for traffic flows are made based on feedback about buffer states and channels. Note that the notion of “channel” is very broad in our model and can represent an abstraction of not only the physical channel but also MAC and PHY layer performance that can be acquired at the PDCP level.

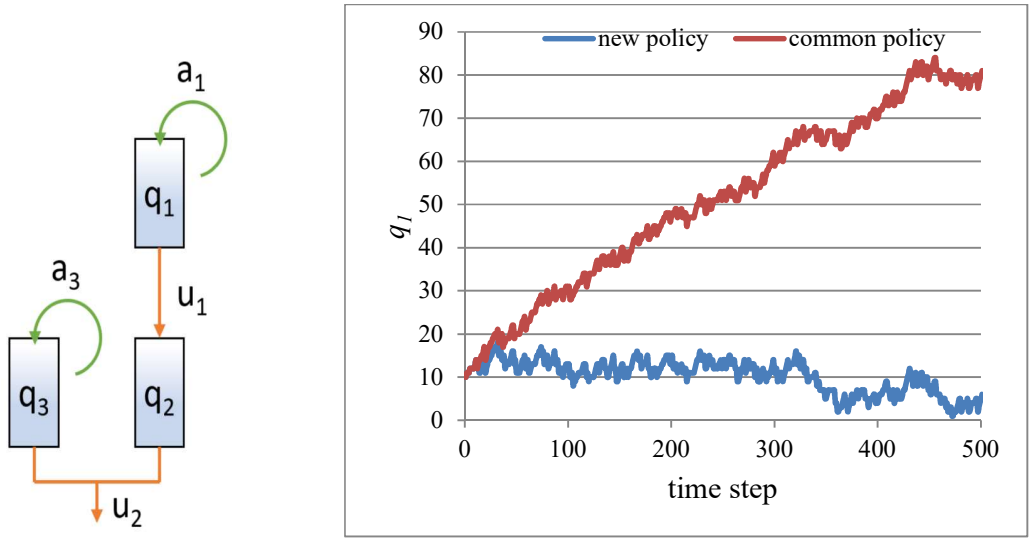


Figure 3-9: Buffer state evolution (right panel) of node 1 (q_1) for an exemplary network (left panel), with conventional and new routing policies.

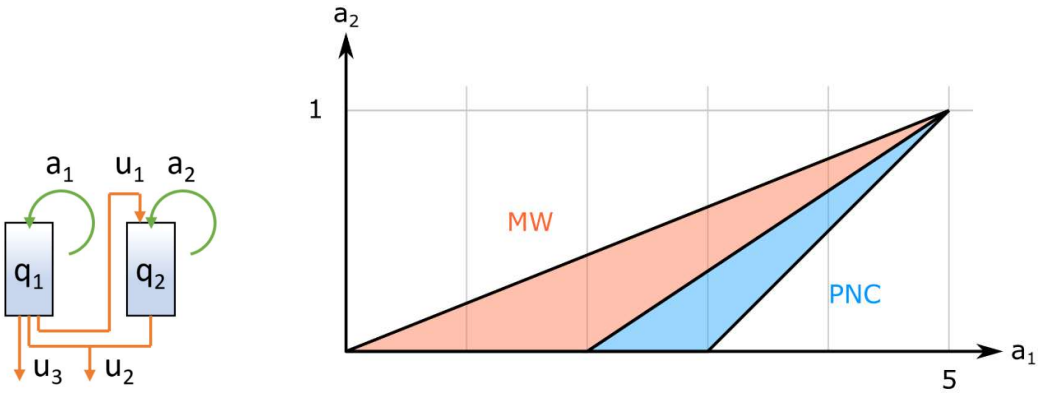


Figure 3-10: Additional stability region (blue area, right panel) of the novel policy compared to old ones (red) for a specific network model (left panel)

3.2.7 Efficient CQI Scheduling

In this section, we address the problem of joint feedback reporting and scheduling in a single cell OFDMA downlink network operating in FDD mode. In order to perform opportunistic scheduling, the BS has to acquire the CQIs from the users. A common method to acquire these CQIs is to schedule the users for CQI reporting. Since reporting the CQIs consumes an important part of the uplink resources, especially for high number of users, only a subset of the CQIs can be reported. This has an impact on the performance of the opportunistic scheduling in the downlink since the scheduling strategies depend on the CQI knowledge for each channel (i.e. RB) and user. In this regard, different approaches are proposed to reduce the feedback load, e.g. [OY13]. While in the literature, the traffic pattern has been ignored in the proposed feedback reporting strategies, the impact of the dynamic traffic is taken into account in this work. In particular, the queueing stability of the users are involved in the feedback strategy. The interaction between the feedback strategies and the queueing stability can be seen easily as the feedback directly impacts the scheduling (and therefore the system stability) since a user cannot be scheduled unless its CQIs are available at the BS. In realistic scenarios, it is of interest to develop a joint feedback and scheduling strategy that takes into account the traffic pattern and the limited feedback (partial and delayed CQI knowledge). Even though the limited feedback has been considered in the literature,

e.g. [OY13], the joint impact of partial and delayed CQI knowledge on the system stability has not been considered so far. In this work, both limited feedback and delayed CQI knowledge are accounted for and their impact on the stability region is characterized. We also provide an algorithm that uses exactly L feedback resources, where L is the number of RBs. The particularity of our algorithm is that the feedback decision is done at the users' side. Our algorithm is optimal in the sense that there is no other solutions that achieves better stability performance under the assumption that only L CQIs can be reported at each time slot. The feedback in this algorithm is done by using a contention-based procedure in continuous time. We also develop an implementation of the algorithm to deal with the case of discrete-time contention. Our approach also takes advantage of the local CSI knowledge of the users in order to achieve better gains. The result in this section is of interest as it shows that if the feedback decisison is done at the user side, the performance will be better than any centralized feedback decision done at the BS. In 3GPP NR, the feedback and scheduling decision is decided by the gNB. The work shows therefore that the current 3GPP standard can be improved by letting the users feeding back the CQIs using a contention-based uplink channel, provided that the decision is performed in a appropriate way. This is mainly due to the fact that a decision at the user side can take advantage of the instantaneous local CSI knowledge by the users. In this section, we provide a brief description of the work. For more details one can refer to [DAD+17].

System Model and Proposed Solution

We consider a FDD cellular wireless network, with one single-antenna BS, N single-antenna mobile users and L RBs. The packets to be transmitted to the users are stored in N separate queues at the BS. Let $q_i(t)$ denote the length of queue i at the beginning of time-slot t. The state of a user on a RB represents the bit rate such that the packets are succesfully received. We assume that each state can take K possible values, which means that the instantaneous rate on each RB can take K possible values (i.e. K Modulation and Coding Schemes).

Furthermore, we consider that the rates vary from one time slot to another due to channel fading, which is modeled here as a *channel convergent* model [Nee03]. Note that this general model is widely used to model a Markovian channel since in general the time varying channel is not necessarily independent and identically distributed (i.i.d) over slots. We consider a realistic context where the CQI knoweldge is delayed and the feedback is limited. This implies that the feedback decision is made d slots before the transmission of the data.

The objective here is to develop a feedback and scheduling strategy that stabilizes the queues of the users whenever it is possible. This means that the feedback and scheduling decision must achieve the stability region of the considered system model. Recall that the stability region is defined as the set of vectors of mean arrival rates for which the queues stay strongly stable.

In the following, we will provide a brief description of the proposed algorithm. More details can be found in Appendix 6.3. In the description, we will focus on the case where the contention is done in continuous time.

1. Queue lengths broadcast every T_b slots:

Every T_b time-slots, the BS broadcasts the queue lengths of all users.

2. *Feedback and scheduling decisions at time-slot t-d:*

This is done by letting the users contend among each other as follows: for each RB, each user waits until a predefined time (one can refer to Appendix 6.3 for more details on how to obtain this time) and then broadcasts a signal of negligible duration to end the contention procedure for the considered RB. The corresponding user reports its CQI. Then, the contention of another RB gets started.

3. *Transmission at time-slot t:*

At the end of the contention period of all RBs, the BS has the CQI of each RB. The user selected to report its CQI of RB j. will be scheduled for data transmission on this RB.

We have evaluated the performance of the proposed algorithm and find the performance gap between our policy and the ideal system (i.e. the system in which the CQIs for all users and RBs are available at each time at the BS at no cost). As mentioned earlier, we have implemented two versions of our algorithm. In the first one, denoted by FSA, the contention is done in continuous time while in the second one the contention is performed in discrete time. An example of the obtained results is provided in Figure 3-11. In this figure, we can see the impact of the amount of feedback, denoted by F, on the second version on the algortihm. It is worth mentioning that in the first version of the algorithm (i.e. FSA) the amount of feedback is always equal to the number of RBs and hence the performance does not change with F (F>L). A description of the theoretical performance analysis is provided in Appendix 6.3. More details on this work can be found in [DAD+17]. As future work, other feedback strategies and more performance analysis schemes will be investigated.

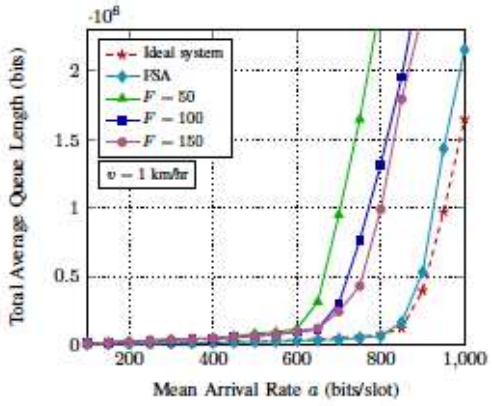


Figure 3-11 Total average queue length vs. mean arrival rate a. N=30 users and L=30 RBs.

3.3 Signalling and Control Plane Optimizations

In this section, different methods for signalling and control plane optimizations are explored in order to ensure basic connectivity in challenging environments such as Megacities and Underserved Areas, by leveraging C-RAN and virtualization capabilities.

Firstly, a mechanism to efficiently perform channel estimation for C-RAN deployments is presented in order to improve user experience in dense scenarios. Secondly, a method to decouple control and data plane signalling is analyzed by using multiple associations in order to reduce low-latency two-way communication in TDD cellular systems. Finally, a method to perform device virtualization is presented with the aid to simplify physical device by performing all connectivity tasks at the network edge.

3.3.1 Signalling optimization for Channel Estimation

The flexibility offered by the (potential) C-RAN architectures considered by 3GPP (see Section 2.2) along with the decoupling of control and data planes, promises improved user experience under the ideal scenario of densely deployed RRHs with CU and/or DUs having knowledge of the channel quality among all possible links (involving RRH-UE pairs) in the system. In that case, the system can, in principle, obtain optimal decisions in terms of, e.g., UE-to-RRH(s) association, transmission/cooperation modes, scheduling of users, and handovers.

However, discussions in 3GPP regarding the, so-called, lower layer split of functionalities are still ongoing (see Section 2.2). There is thus no provision in the first phase of NR for efficient signalling towards achieving the full premise of advanced resource allocation and/or transmission

techniques that are possible in a C-RAN setting. This, in turn, implies that C-RAN deployments in NR are (currently) treated as conventional (i.e., D-RAN) deployments with sophisticated techniques taking advantage of the C-RAN architecture implemented proprietary. However, in a setting with a massive number of densely deployed RRHs and users, as envisioned for dense urban areas, any (proprietary) attempt to achieve the full potential of C-RAN must face the challenge of excessive signalling overhead required for obtaining global channel state information (CSI), under the orthogonal-signalling-based CSI-RS framework currently considered in NR. This is because the number of links in the system is proportional to the square of nodes in the system requiring a proportional number of dedicated resources (e.g., REs) in order to accommodate the transmission of orthogonal CSI-RS from each node in the system.

Clearly, channel estimation procedures, involving both (a) RS design and (b) estimation algorithm should be devised towards minimizing the signalling overhead with tolerable degradation of channel estimation quality. Towards this end, an important observation is that, in a scenario where the C-RAN is deployed over a large geographical area, as in the use cases of smart cities or airports, most of the RRHs will not have strong links to any given UE (due to path losses). Taking this into account, we propose the following channel estimation procedure for the downlink C-RAN channel estimation: all RRHs transmit their unique CSI-RS on the same, but of limited number, N_p, radio resources (e.g., REs). At the UE side, the superposition of the non-orthogonal RSs is received (see Figure 3-12) and the UE task is to obtain accurate global CSI based on this. Clearly, a channel estimation algorithm assuming orthogonal RS will fail due to interference. One approach to obtain accurate CSI in this setting is to exploit advanced algorithms from the field of compressive sensing [WBS+15], [FZY16]. This approach is the topic of extensive research in D-RAN scenarios (see IR.4.1, Chapter 3) and is applied here in the C-RAN setting, exploiting the power-domain sparsity of the C-RAN topology as viewed by any random UE in the system. The question here is: "what is the trade-off between signalling overhead and CSI estimation quality?".

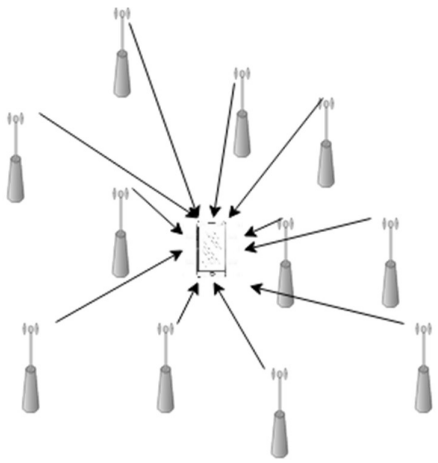


Figure 3-12 Illustration of considered signalling scheme for downlink CSI acquisition in C-RAN. All RRHs transmit their unique but short-length training sequence on the same (signalling) resources and their superposition is received by an arbitrary UE in the system.

As a first step towards answering this question, the, so-called, oracle estimator analysis was pursued that provides performance bounds for most of the commonly used compressive sensing algorithms. We showed in the recent publication [SW18] that for a scenario where the RRHs are randomly deployed on the plane, all links experience independent, identically distributed Rayleigh fading, and the propagation path loss exponent equals $\alpha > 2$ for all links, the mean square error of the channel estimation for an arbitrary UE cannot be smaller than the value

$$(N_p - 1)\left(\frac{\pi\alpha\Gamma(1+2/\alpha)\lambda}{(\alpha-2)(N_p-1)}\right)^{\alpha/2},$$

where $\Gamma(\cdot)$ is the Gamma function. Here, λ denotes the average number of RRHs per unit area (density of RRHs). This simple analytical formula clearly demonstrates how the signalling overhead (represented by the number of signalling resources N_p) affects channel estimation performance as well as the effect of the propagation conditions. As expected, increasing N_p improves MSE, however, this improvement is greater when increasing α. In addition, for the same N_p, larger α results in better MSE, whereas for α close to 2 the MSE becomes unacceptably large. This shows that there is a *fundamental physical limit* in reducing the signalling overhead in C-RAN, with high path loss exponents, experienced, e.g., in dense urban environments, being beneficial. Numerical examples in [SW18] indicated that for path loss exponents greater than 4, high quality global CSI can be acquired with a signalling overhead that is less than half than the one that would be required by conventional, i.e., based on orthogonal sequences, methods currently considered in NR.

Future work on this topic will consider the application of these theoretical insights to the design of the NR CSI-RS framework, towards enabling the full potential of C-RAN deployments, i.e., network MIMO operation.

3.3.2 Separate control information and data via multiple association

A reliable communication link is established through a two-way communication protocol, where the receiver always acknowledges (control information) the reception of the transmitter's packet (data). This brings forward the need to have two-way communication links where each device can quickly switch between transmission and reception. The obvious way to achieve low latency is to use full duplex transceivers using Frequency Division Duplex (FDD) systems. Contrary to this, many ongoing efforts are currently favouring Time Division Duplex (TDD) operation to reduce transceiver cost, increase spectrum usage efficiency, take advantage of channel reciprocity and to be capable to adapt to time-varying uplink/downlink traffic asymmetries [PBF+17]. However, the frame-based structure of TDD is not aligned with the requirement for latency reduction due to the long time it takes to switch between uplink and downlink. Based on these observations, we conclude that there is a need to design low latency two-way communication solutions that can work with terminals that operate in TDD.

In TDD cellular systems, the minimization of the latency experienced by a two-way communication transaction is constrained by the frame duration that the half duplex base station stays in the downlink and uplink directions as shown in Figure 3-13(a). The direct approach is to decrease the time period before shifting from uplink (downlink) to downlink (uplink) direction, in order to facilitate fast interaction between the communicating devices/nodes. This fast switching between uplink and downlink comes at the cost of more complex transceivers both at the device and base station, due to the need to perform faster and more frequent channel estimation, as well as additional signalling overhead, due to the resource assignment.

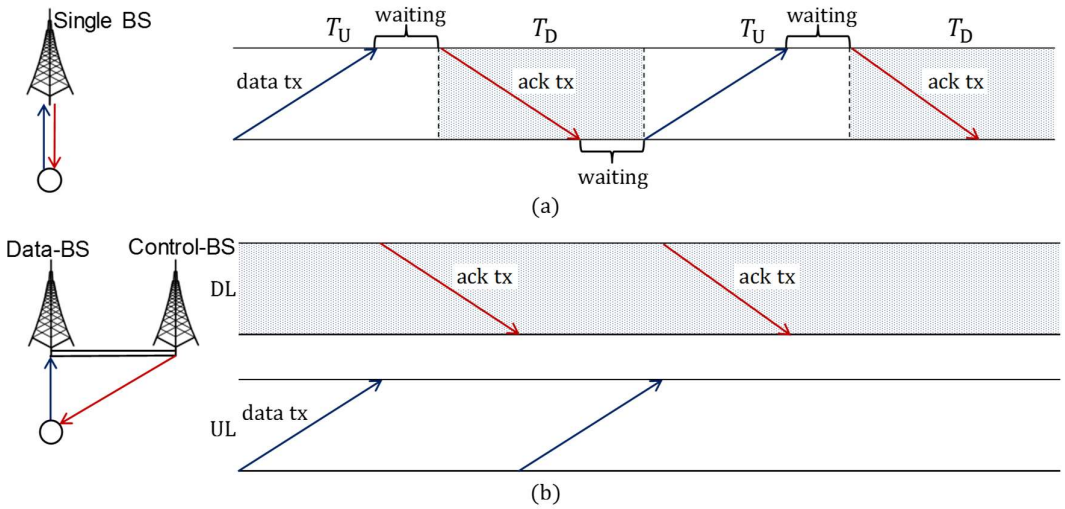


Figure 3-13: Illustration of separate control information and data via multiple association.

We propose a radically different method for enabling low-latency two-way communication in frame-based TDD cellular systems. We propose to use two half-duplex base stations, where the first base station takes the control information and the second the data (or vice versa). As a motivation example on how this approach can satisfy the individual low-latency requirements, consider the example in Figure 3-13(a) with two devices where each should receive a reply from their original transmission within 2 timeslots. The TDD structure of the baseline scheme in Figure 3-13(a) is not capable of avoiding additional waiting times; while, as shown in Figure 3-13(b), the proposed scheme can support the desired latency, as it allows each device to switch between uplink and downlink directions without requiring the network infrastructure to do so. This of course assumes that the processing time within the device to switch between UL and DL directions is lower than the residual time remaining in the traditional cellular TDD to switch between UL and DL (and vice-versa).

The proposed scheme can be applied both in a traditional and in a C-RAN architecture. In the traditional architecture, the coordination between the half-duplex base stations can be accomplished through the X2 interface; while in a C-RAN architecture this coordination is implicit at the C-RAN's base-band unit (BBU). The interference resulting from the downlink to the uplink half-duplex base stations can be mitigated by: (i) Taking advantage of spatial pre-computing, beamforming and full-dimension MIMO to steer the downlink interfering beam from the uplink base station; and/or (ii) take advantage of the X2 interface to exchange the necessary information to cancel any residual wireless interference, with the goal of improving the reliability of the uplink reception in a low latency traffic setting.

The proposed scheme builds upon already on-going efforts in 3GPP such as device multi-connectivity, decoupled uplink and downlink access, cooperation between base stations and dynamic TDD. Furthermore, it enables the network to continue its evolution towards a device centric architecture and enable the joint design and scheduling of low latency two-way communications.

In TDD systems, the proposed scheme, where the Half-Duplex Base Stations (HDBSs) are connected via an X2 interface (Figure 3-13(b)), can achieve lower latencies than the single TDD baseline (Figure 3-13(a)). The baseline scheme is composed by a half-duplex device and a HDBS as shown in Figure 3-13(a). The simplest realization of the proposed scheme is depicted in Figure 3-13(b). Each transceiver on the picture, both at the UE and at the infrastructure, is half-duplex and operates in a TDD mode. The detailed performance evaluation can be found in Appendix 6.7.

3.3.3 Virtualization and RAN functional split aspects

Both "Megacities" and "Underserved Areas" demand suitable optimizations of control plane functionalities to enrich their capabilities in challenging environments. "Underserved Areas" can pose very challenging operational conditions in real deployments, because of e.g. expensive backhaul connections (satellite, microwave links…), low reliability of the physical assets (infrastructure, energy supply, etc.) or the presence of natural disasters. "Megacities" scenarios can also pose challenging conditions as per the need to tailor the radio access functionality to the specifics of the services being offered through network slicing.

In both cases, C-RAN (section 2.2) represents a smart lever to tailor deployments as per the needs of network slicing. ONE5G will investigate techniques to cope with such challenging conditions in both scenarios by means of C-RAN, RAN virtualization, and RAN functional split.

Referring to the split points studied in 3GPP RAN3 (

Figure 2-5), it is apparent that each split option gives rise to different transport network needs (mostly in terms of throughput and latency), and in turn allows different advanced RAN functionalities like e.g. coordinated scheduling, coordinated beamforming or joint transmission etc.

Beyond the traditional RAN coordination capabilities, the need to offer ultra-dense connectivity (in the "Megacities" scenario) can greatly benefit from virtualization of the user device, in connection with C-RAN and MEC capabilities. To this end, the physical user device can be permanently connected to a Device Virtualization Server (DVS), which resides at the RAN with the aid of MEC techniques. Multiple device functionalities can be flexibly executed thanks to the DVS node, which implements all user interface and connectivity procedures involving the user device (Figure 3-14).

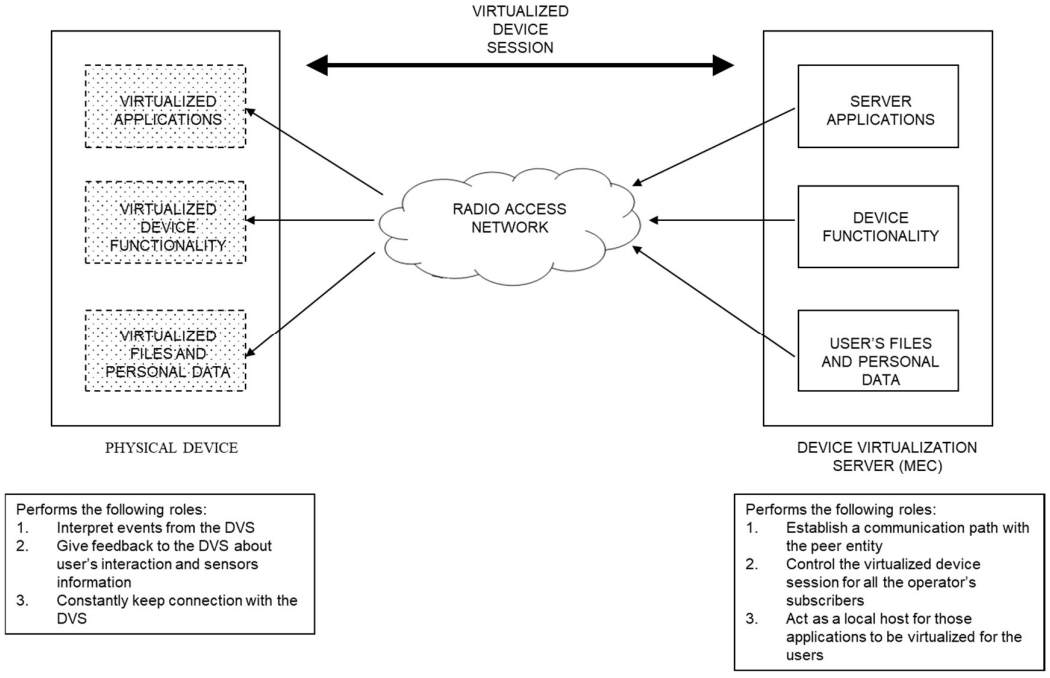


Figure 3-14: Illustration of the Device Virtualization Server, leveraging MEC techniques, that hosts the virtualized device functionalities

The aid is that mobile users can get rid of any actual constraints imposed by physical device by virtualizing its functionalities at the network side. Hence, the users get the impression of a much more capable device since the functionalities are provided by the network, independent of the actual device capabilities.

The DVS node has twofold function: deliver virtualized device sessions to the users in the mobile network, and centralize all connectivity tasks between the users and any other entities in the network. Moreover, it would gather a portfolio of third-party applications that will be used by the users so that all subscribers can benefit from them without the need to install, upgrade, or download any user data as everything would reside in the cloud, and accessed by the user through the DVS.

For low-latency, permanent connection to the physical device is assumed to be always present, thanks to the ultra-high density of access nodes. UE applications and connectivity protocols are managed by the DVS in such a way that the physical device would be greatly simplified, while retaining the full characteristics and functionalities of the user device. The physical user device will comprise in turn three parts:

- an **operating system (OS)** containing a local boot system and basic functionalities to run some applications locally when the user is out of coverage.
- a set of **HW resources** with all kinds of sensors and interfaces needed to interact with the network,
- and, a **communications module** to address the requests and events received to/from the network.

Specific C-RAN architectures and functional split points have been therefore explored to support this use case concluding that it is more suitable for C-RAN deployments with low-layer split where all the radio network functions above low-PHY are centralized at the CU. Low-layer split allows improving user experience by means of complex centralized coordination techniques such as CoMP, centralized scheduling, coordinated beamforming, etc.

The proposed architecture is illustrated in Figure 3-15 where the DVS node is placed at CU to take advantage of RAN parameters available through RNIS (Radio Network Information Service) APIs, hence making the DVS aware of the actual radio conditions at the nodes. Nevertheless, DVS can eventually be moved closer to the user, at DU, for those applications which require very low-latency requirements.

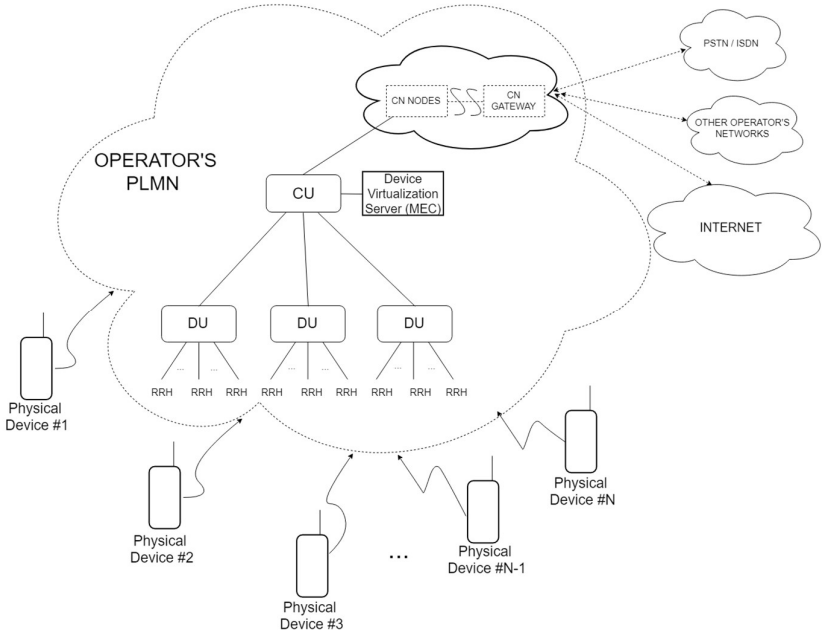


Figure 3-15: Device virtualization architecture.

4 Multi-link Management for Improved E2E Performance

This chapter presents the work carried out as part of the task focusing on multi-link management for improved E2E performance. Specifically, the included solutions cover topics such as dynamic multi- connectivity (MC), spectrum management, mobility optimization and performance optimization for D2D schemes. For each topic, an introduction and the status in 3GPP standards are presented firstly. Finally, the proposed solutions are described.

4.1 Dynamic Multi-Connectivity

In this section, the topic of dynamic Multi-Connectivity (MC) will be introduced, along with a first set of formulated problems and solution frameworks towards improving network performance, with respect to, e.g., the experienced data rate, latency and reliability. More specifically, management and self-optimization mechanisms will be discussed, focusing on deciding upon which users should operate by means of single/multiple connectivity. Referring to multi-link users, the performance-driven applicability of the operation mode (i.e., data duplication/aggregation) for each link will be discussed, along with possible DL/UL decoupling. Throughout the section, the flexibility of system design will be emphasized, driven by system dynamics, where, for example, a user's single-link/multi-link connectivity is expected to vary during an active session.

Introduction to Multi-Connectivity

Multi-Connectivity (MC) is a radio network feature which has recently attracted a lot of attention, due to the advantages of aggregating radio resources with regards to satisfying several (and dissimilar) performance requirements of current and emerging service types (eMBB, URLLC, mMTC). Use of MC is particularly relevant for the ONE5G megacity scenarios [D21]. Contrary to single connectivity, where a device is always connected to a single point of transmission/reception (e.g., an Access Point (AP)) with the quality of service depending on the quality of this single link, in MC, multiple APs can simultaneously configure radio resources to a given terminal, introducing link diversity. A direct benefit lies, for example, in the case of a link failure, where the device will still be connected to the rest of the nodes, thus, improving connection reliability. It should be noted that, as future network deployments are envisioned to be dense, such density can be translated into the existence of multiple strong links to be exploited by a device running a demanding wireless application, however, technical solutions are still needed in order to deal with the increased interference, as well as to reduce the processing complexity at the terminal side.

Focusing on 5G networks, different MC variants exist, which can be first divided into inter-frequency and intra-frequency solutions. As explained in Figure 4-1, both inter-frequency and intra-frequency MC solutions can be further sub-divided into inter-site (which may also refer to as Multi-Node Connectivity (MNC)) and intra-site solutions. More specifically, Carrier Aggregation (CA) and inter-site MC correspond to intra-site and inter-site variants of inter-frequency MC, respectively, while Joint Transmission/ Dynamic Point Selection Coordinated Multipoint (JT/DPS-CoMP) and Multi-Flow, correspond to intra-site and inter-site variants of intra-frequency MC, respectively. Each of these MC variants provides performance-based benefits, as further explained in the figure.

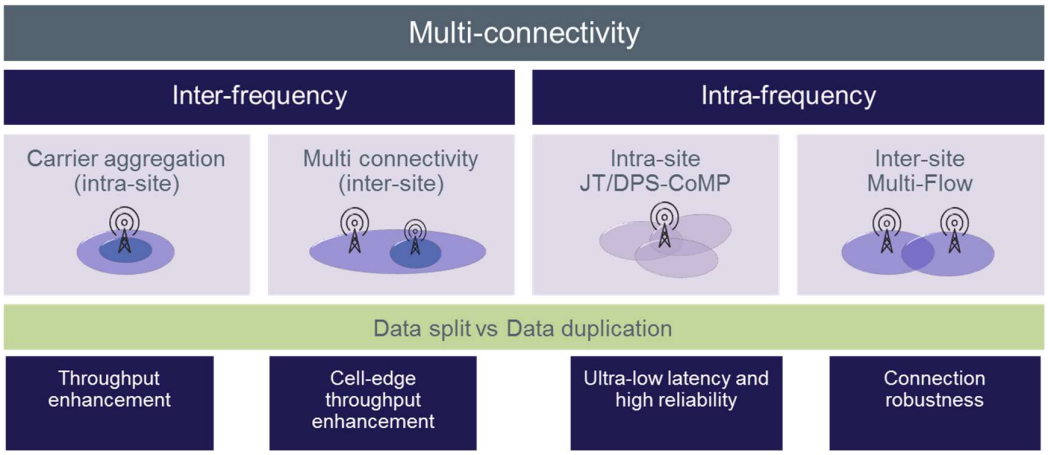


Figure 4-1: Variants of MC solutions

Regarding the treatment of data by the various nodes providing connectivity to a user device, two distinct options currently are considered: data split and data duplication. The main advantage of the first data management method is that, by means of splitting and distributing the data among the nodes, the user throughput is expected to increase, whereas, by exploiting data duplication, i.e., all nodes transmit the same data, communication reliability can be enhanced. MC for data duplication, which will be discussed later on in this section is not a very new concept from a PHY standpoint, since, e.g., in CoMP JT multiple sites transmit the same data.

MC status in 3GPP

NR inherits the Dual Connectivity (DC) design from the E-UTRA DC functionality specified in 3GPP Rel-12/13 [36.300]. In short, LTE DC extends Carrier Aggregation (CA) operations to non-collocated nodes connected via non-ideal backhaul, allowing a UE configured with split bearer to receive (and likewise send) data from (to) two distinct eNBs simultaneously for the given bearer. The data for such bearer is split at the PDCP transmitter, transmitted via the two radio paths and aggregated at the PDCP receiver, resulting in an end-user throughput boosting rather suitable for MBB services [RPW+16].

To enable a faster introduction of 5G NR, the first NR deployments will be non-standalone and complementary to LTE, reusing the existing evolved packet core (EPC). For that purpose, 3GPP has generalized the LTE DC design for the support of Multi-RAT Dual Connectivity (MR-DC), i.e. DC between NR and LTE, which is limited to two nodes both for the downlink and uplink, in the "early drop" of Release-15 of 5G NR [37.340].

Although Multi-RAT DC had been already introduced in 3GPP Rel-13 between LTE and WLAN in the form of LTE-WLAN Aggregation (LWA) [36.300] [LLR+18], the new design is tailored to the cases that any involved RAT in the dual connectivity operations is of a 3GPP standard. This removes the specific challenges due to the presence of a non-3GPP RAT.

The most prominent architecture within the MR-DC family is E-UTRA-NR DC (EN-DC). In the EN-DC approach, the UE is connected to an LTE eNB, i.e. Master eNB (MeNB) and a NR gNB, i.e. Secondary gNB (SgNB), which are interconnected via a non-ideal backhaul X2 interface. As next step, NR-NR DC, i.e. DC between two standalone gNBs connected to the 5G core operating without the assistance of the LTE eNB, and interfaced with the new Xn interface, is expected to inherit the MR-DC design and be supported as part of Release-15 "late drop".

Furthermore, NR DC (denoting DC between two nodes of which at least one is a NR gNB) extends LTE DC further towards a solution to boost reliability for URLLC [38.300] [37.340]. This is achieved by using data duplication at the PDCP layer rather than data split, in which the *same*

data packet (i.e. PDCP PDU) is independently transmitted over to the same UE through multiple distinct gNBs, increasing the likelihood of successful reception.

The NR PDCP specification in Rel-15 allows to map the duplicates to two RLC instances which can belong to the *same* Cell Group (CG) (i.e. a given UE has only one set of serving cells) to support CA, or to different CGs in the case of DC [38.323]. The duplicated packets would then be served either through distinct component carriers (in CA scenarios) or district nodes (in DC scenarios). Therefore, duplication is only supported if the number of available frequency carriers is larger than one. It is noteworthy that in CA where the same CG is used, specific restrictions at the MAC are placed to guarantee that the two duplicated packets are not transmitted on the same carrier, which would vanish the duplication benefits [38.321].

Looking ahead, on top of the basic functionality, Rel-16 is expected primarily to enable the extension of DC to consider more than two cells and smoothen NR DC operations during mobility events.

4.1.1 Flexible cell connectivity based on multi-service requirements

Among the envisioned adverts of future cellular systems including 5G and beyond, multi-service requirement stands out as a key element of such systems. Motivated by the ever-increasing network traffic, a paradigm shift from the conventional homogeneous networks to multi-tier heterogeneous networks (HetNets) is envisaged. HetNets are composed of macro Base Stations (BSs) overlaid with lower tier BSs (BSs with lower capabilities and denser spatial deployment per unit area) such as pico, femto, and relay BSs. Macro BSs with their high transmit power provide large coverage areas, while lower tier BSs serve users in coverage holes and hot spots.

Conventionally, in state-of-the-art cellular systems, the downlink (DL) received signal power determines the serving cell to which the UE will be connected. Such a criterion, known as the maximum Reference Signal Received Power (RSRP) association criterion, is optimum for homogeneous networks, where the transmit power is nearly equal among different BSs. Nevertheless, for multi-tier HetNets incorporating different types of BSs, a large power disparity in the network exists, hence, leading to imbalanced loads among the multi-tier BSs [LBY+15]. An illustrative example of cell connectivity based on the maximum RSRP and nearest BS criterion is shown in Figure 4-2. It is observed that the large disparities of the coverage regions in the download, resulting from the transmit power disparity between the two tiers, lead to a load imbalance regarding the associated users. Such a heterogeneity of coverage is non-existent in the uplink, since the transmit power disparity in the uplink (UL) is smaller compared to DL. In the mentioned figure, the green squares represent the users

To alleviate the drawbacks of the large disparity between DL and UL coverage areas (i.e., load imbalance and non-effective resource allocation), downlink and uplink association decoupling has been proposed as a disruptive solution for an enhanced network performance [HFM+14]; mainly giving the users the flexibility in the uplink to associate with the BS that provides the minimum pathloss. Such a solution is shown to provide a higher achievable throughput compared to the Cell Range Extension (CRE) solution proposed in 3GPP Rel-12 [FJH+16].

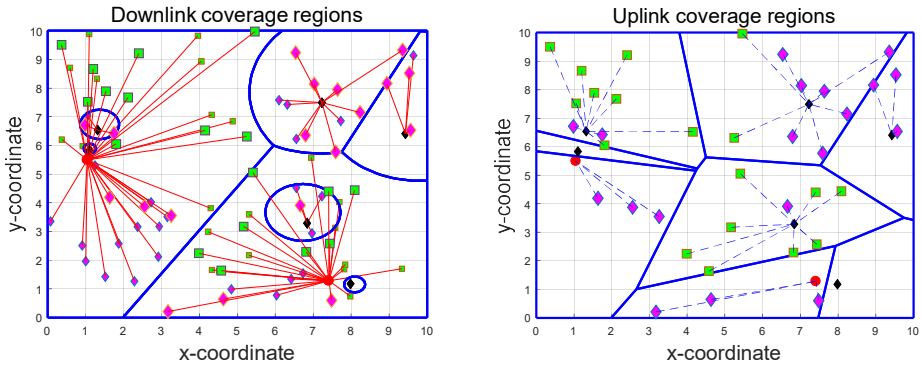


Figure 4-2: Coverage areas (left - DL and right - UL) for a two-tier HetNet consisting of users (green squares) and macro eNBs (red circles) overlaid with small BSs (black rhombuses) for a specific network deployment & channel instantiation. Solid red lines represent connectivity achieved by applying the maximum RSRP rule, while blue dashed lines represent connectivity under the minimum path-loss criterion. The green squares and magenta rhombuses represent the coupled and decoupled users, respectively.

Throughout this work, we aim to study the flexibility of the cell association rule via allowing users to attach to different BSs in the DL and UL (i.e. downlink/uplink decoupling). Such a flexible association is considered a key enabler to meet the diverse QoS requirements envisioned for 5G. Moreover, a main motivation for exploiting the decoupled association is the increased spatial heterogeneity in the network. To clarify the concept of decoupled association, assume that the transmit powers for a two tier network, consisting of macro and small BSs, are denoted by P_M and P_S, respectively, while, a randomly selected UE will connect to the macro BS if, $P_M||x_i - y||^{-\alpha} > P_S||z_i - y||^{-\alpha}$ where $x_i, z_i \& y$ represent the macro BS, small BS and UE locations, respectively. Naturally, an association to the small BS will happen otherwise. It is worth noting that the association procedure (typically done based on the RSRP) is obtained by averaging over the received signals with respect to the channel fading [KPL15].

In order to observe the effect of different network parameters (spatial density and transmit power disparity), extensive simulations were conducted to observe the association events, as shown in Figure 4-3. First, regarding the effect of the increased small BS density on the association probability (i.e. left figure), as the ratio of small BSs to macro BSs increases, one would expect the macro BS coupled association to decrease and that of the small BS to increase, which is verified by the mentioned figure. Additionally, it is observed that the decoupling probability increases till a drop point after which it decreases at the expense of increased probability of the coupled association to the small BSs. Such a behavior is attributed to the closer small BSs in the vicinity of the randomly selected user, thus, leading to a coupled access instead of a decoupled one. On the other hand, an important aspect is the relation between the decoupling probability and the network resource disparity (i.e. transmit power disparity). In the right part of Figure 4-3, we plot the probability of decoupled association as a function of the small BS spatial density for two values of the inter-tier transmit power disparity. One can conclude that, as the transmit power disparity increases, the probability of decoupled association increases, due to the increased dissimilarity of the coverage regions of the two tiers, which can be better visualized in Figure 4-2.

Finally, future work will focus on the development of a joint, QoS-aware cell association and resource allocation framework, given the challenge that the available resources at the BS (i.e., Physical Resource Blocks - PRBs), along with the imposed QoS constraints from the users, highly affect the choice of the serving BS itself.

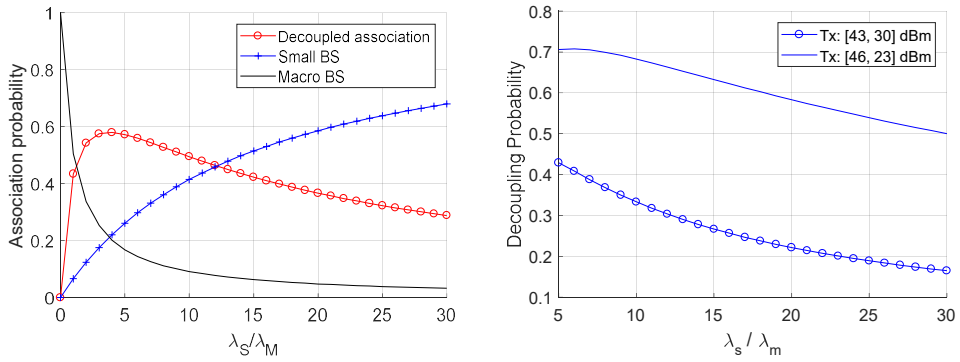


Figure 4-3: (Left) Association probabilities for decoupled and coupled association with increasing spatial density of small BS. (Right) Decoupling probability as a function of the increasing spatial density for two variants of BSs transmit power disparity.

4.1.2 Low-latency two-way communication in a cellular network

The network protocol must work properly for smooth data transmission and control information must be delivered in a timely manner. For example, transmitting data in one way can be acknowledged by the opposite direction response. These information exchanges form a two-way communication. It is important to give a fast control message response because the sender is waiting for another action until the response comes back. To achieve this, we can adopt multi-connectivity.

Network densification and multiple BS utilization can reduce latency of two-way communication. The decrease in BS-user association distance due to densification mitigates the propagation loss, which is important for the most severely affected users because cell association is based on the maximum RSRP criterion for 3GPP NR. In the noise-limited regime, where aggregate interference is negligible compared to noise, network densification increases the desired signal power and improves the reliability. For the interference-limited regime, the short propagation distances due to network densification increase not only the desired signal power but also the interference that may be generated by numerous neighboring BSs. Nevertheless, the desired signal power increase dominates the increase of interference due to the path-loss which follows a power-law. To the end, it could be possible to increase signal-to-interference-plus-noise ratio (SINR) for all users [PKZ16].

Network densification also leads to resource reuse and increases per-user resource allocation. This resource increment can be directly utilized for latency reduction. Alternatively, it can be dedicated to diversity for reliability enhancement. Finally, network densification makes BSs more likely to have a few or even no associated users within their coverage, especially in ultra-dense network setups where the BS density exceeds user density. Such user-void BSs are expected to be in an idle state, not sending data signals for energy-efficiency, but may provide extra associations for the URLLC users. This, however, increases the downlink interference from the awakened BSs, which can be mitigated by cooperation among neighboring BSs.

Consider two neighboring BSs that are interconnected through a high-speed backhaul, thanks to their short inter-BS distance after densification. In order to illustrate the concept, assume that the network features two types of users: low latency users (LLUs) and latency-tolerant users (LTUs). By exchanging data signals and association information, these two BSs can serve their users concurrently without incurring interference. This can be achieved by utilizing interference cancellation or prioritizing the transmission of low-latency user. The detailed multi-BS cooperation methods is found in Appendix 6.6.

Figure 4-4 shows its effectiveness in average latency reduction. If the portion of LTUs is high, the performance of LLU is improved using the proposed scheme. If all users in the network are

LLUs, the network resource should be increased, in that BS densification is necessary to achieve certain latency.

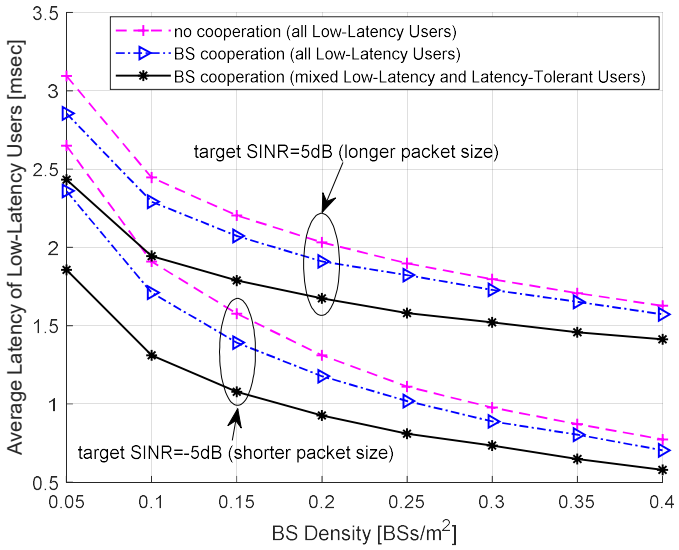


Figure 4-4: Average latency reduction from BS densification (x-axis) and extra associations with BS cooperation (blue triangle and black star) for user density 0.1.

4.1.3 Component carrier management

5G is expected to support users holding different services categories [ITU2015]: eMBB, URLLC and mMTC, each of which presents different traffic profiles and performance requirements. The objective of this work is to make use of multi-connectivity, aggregating radio resources in an efficient and flexible manner, so as to meet such requirements.

Several techniques for component carrier management have been proposed, each of them following different criteria [WAN10, LVM17]. The most immediate research line could be the load balancing among component carriers. In the same way, dual connectivity has been addressed in recent works, showing its advantages and capabilities in different scenarios [ROS16, LEM16], for example, regarding its ability to reduce radio link failures given a fast-moving UE. Finally, some recent works propose addressing multi-connectivity through the 3G concept of active set management [TES16]. However, in these works, only the radio channel conditions are considered as their input for the component carrier management.

The aim of this work is to dynamically assign component carriers from multiple (more than two) nodes (extending dual connectivity) according to the network state (e.g., network load or coverage hole), as well as the service category and context information. In this study, only eMBB and URLLC will be considered (i.e. ONE5G use cases no. 2, 5 and 6 [D21]) and will be managed in a different way considering their different requirements. For URLLC, the reliability will be addressed through data duplication. For eMBB, given its need for higher throughputs, a data aggregation scheme will be followed. To this end, a Component Carrier (CC) manager is proposed to determine the number of carriers to be assigned to a user, as well as the carrier indices and the source nodes, and flow control (e.g., data aggregation or data duplication). That is, using a duplication or aggregation scheme. Different types of inputs are considered, such as: (a) metrics reported by the user, like the RSRP; (b) metrics from the carriers (like their load); (c) metrics from end-to-end information (like throughput or latency) and (d) information from the context (like the user position). Based on these inputs, the CC manager computes a score for each of the available carriers indicating the carrier suitability for a specific user. This score can be computed

by different ways depending on the target criterion (e.g. if a load balancing approach is followed, those CC with a lower load will receive a higher score).

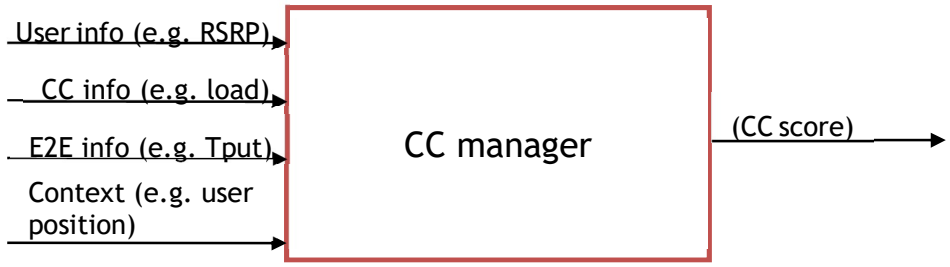


Figure 4-5: Schematic of the CC manager, including its inputs and outputs.

A first version of the CC manager is being implemented using a fuzzy logic approach. This technique is especially suitable to be applied to cellular networks since it allows operators to map their policies to the management functions. This implementation is being carried out on a system-level simulator developed in MATLAB that includes the multi-connectivity feature. Only RSRP values measured by users and information about CC load are considered as inputs in this first version. In the first phase of the process, input values are translated into linguistic terms that represent different situations by means of membership functions. For example, RSRP values could be HIGH or LOW (e.g., an RSRP value would be HIGH if it is higher than -90dBm). Then, the score for each CC is computed by applying a set of IF-THEN rules over the categorized variables that implements a specific policy. In this first version, the rules have been defined to achieve a load-balanced situation among CCs. As a result, the CC manager would assign for a certain user the CCs received with the highest value of RSRP only if the corresponding load is not very high.

In addition, some preliminary analysis has been performed using multi-link single-node simulations with the LTE module of NS3, namely LENA [BAL+17]. We have examined the performance of UDP downlink data traffic from a remote server to a single UE, being the only UE in the system. This UE is located at 60 m from the base station. All the possible combinations of number of CCs (from 2 to 5) and number of PRBs per CC (6, 15, 25, 50, 75, 100) are assessed. The UE is connected to a single eNodeB, which has wired connections with the S/P-GW. The S/P-GW links to the remote server based on a high-speed point-to-point (P2P) connection of 100 Gbps with a delay of 10 ms. End-to-end throughput and delay measurements are shown in Figure 4-6.

We conclude that, as the number of component carriers increases, the throughput increases proportionally and the delay decreases. Further studies will be carried out considering more realistic environments.

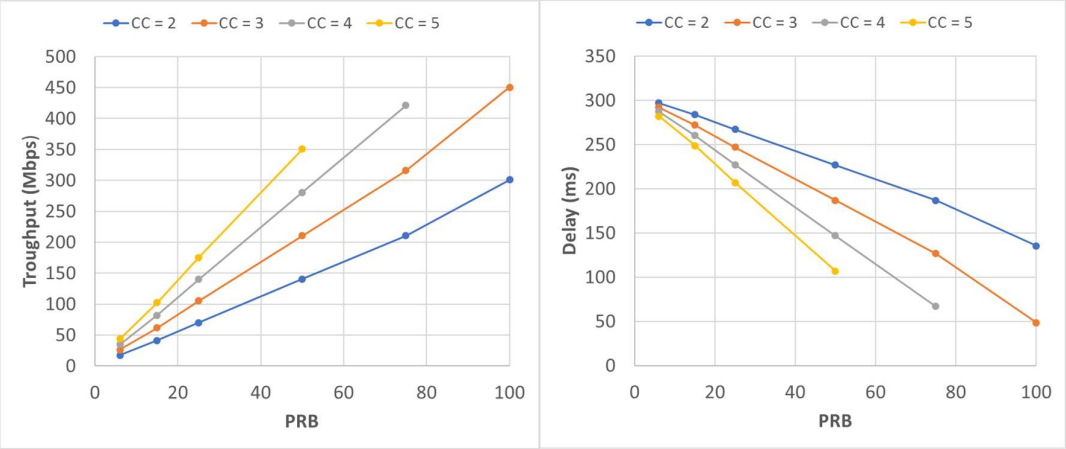


Figure 4-6: Mean E2E (UDP) throughput, (a), and delay, (b), experienced by a UE given different combinations of system bandwidth and number of CCs.

With the aim of improving the services implementation in the simulation tool, TCP has been included in the simulator. User's congestion window is modeled following the TCP version Reno rules. Also, RTT depends on congestion window size.

In order to validate the implemented TCP model, two simulations have been carried out. Tests have been performed in a non-homogeneous scenario with static users. Furthermore, the number of PRBs per CC is 50. In the first test, files are downloaded through TCP while DC is not activated. Then, the obtained results have been compared with the obtained values when DC is activated. User throughput and the number of lost segments are shown in Figure 4-7.

We conclude that, in general, user's throughput is higher when DC is activated. The improvement achieved with respect to the case without DC is more noticeable when users experience worse radio conditions. As for the number of lost segments, the higher values are obtained for users with worse conditions, as expected. In this case, a significant improvement is also achieved when DC is used. Nevertheless, despite its benefits, the usage of MC would entail additional signaling load, both in the air interface and the backhaul, increased processing requirements at the UE and the base station side and the allocation of more resources per user. As future work, more studies will be carried out considering a higher number of links. Additional results will be provided as well. It is worth mentioning that, in collaboration with WP2, the work in this section is proposed for implementation in WP2 system level simulations.

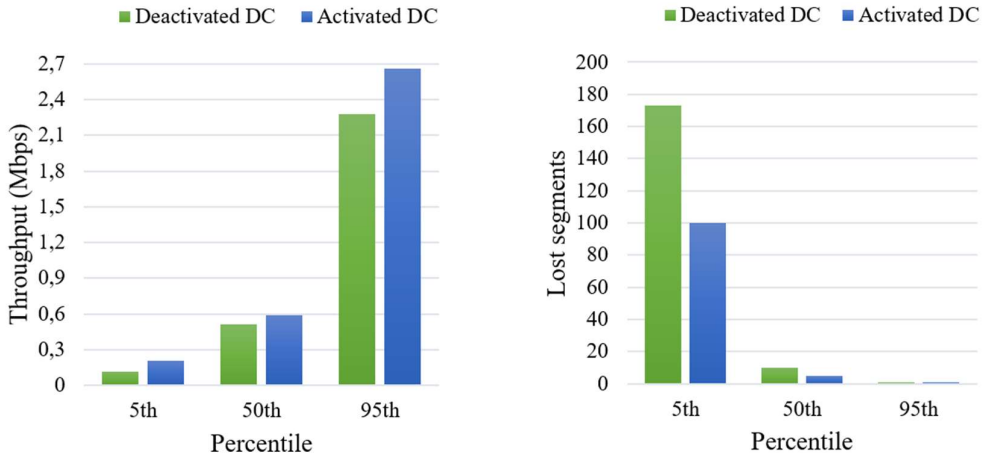


Figure 4-7: 5th, 50th and 95th percentile throughput, (a), and lost segments, (b), experienced by a UE given activated and deactivated DC.

4.1.4 Reliability-oriented multi-connectivity for URLLC

In 5G new radio (NR) standardization activities, DC/MC is proposed as a potential reliability enhancement solution for URLLC applications. Reliability-oriented DC/MC introduces diversity through an additional secondary node in addition to the serving master/anchor node, as schematically presented in Figure 4-8 for the specific case of DC – data duplication through the anchor and a single secondary link. Ideally, the reliability resulting from duplicating across K independent links (each having outage probability p_k) is $1 - \prod_{k=1}^{K} p_k$.

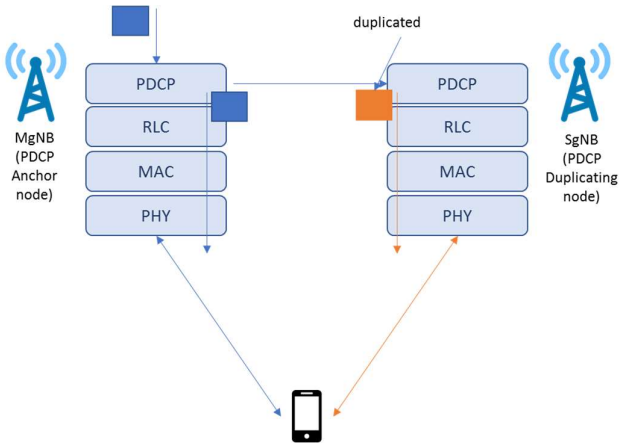


Figure 4-8: Dual connectivity for data duplication.

Challenges in Reliability-Oriented DC/MNC

We aim at addressing some of the main challenges associated with reliability-oriented DC/MC, as further elaborated herein. Firstly, the success of DC/MC in boosting the reliability and/or improving the latency depends on a clever selection of the proper secondary node that offers the best reliability and latency performance to the serving user. However, the reliability for a given UE through a candidate secondary gNB depends on a number of factors such as the signal quality, interference conditions, load at the candidate SgNB. Such information is not readily available at

the anchor gNB (i.e. the node which determines the data duplication policy), and efficient mechanisms to signal this information are needed.

In DC/MC operation, multiple gNBs – termed as the duplication set – *independently* transmit the same PDCP packet. When the PDCP packet is successfully received from a radio link (i.e. from any one node within the duplicating set), copies of that packet received through other links in the duplication set are discarded at the UE, resulting in unnecessary transmissions and additional interference. Hence, there must be mechanisms to avoid such redundant transmissions. Moreover, although the standardized mechanism of keeping the first successfully received PDCP packet makes use of diversity gain, the potential array gain is not exploited. A more efficient mechanism would be to combine the packets arriving from the duplicating nodes to fully utilize the duplicated transmission. However, how to do so remains an open question.

Methodology and preliminary results

Our contribution within this work item is two-fold. Firstly, we propose solutions for the challenges outlined above that are implementable within the 3GPP NR standardization framework. Secondly, reliability-oriented DC/MNC, including the proposed solutions, will be evaluated through system level simulations to verify the achievable reliability gains.

Herein, we present preliminary baseline results of single connectivity in the considered HetNet scenario with a 2 GHz macro layer and a 3.5 GHz pico layer, deployed as a cluster of 4 pico cells per macro area. Each macro area serves on average 30 randomly dropped URLLC users with 2/3 of the users dropped within the pico cluster, where each transmission is operated at a 1% BLER target [PSP+17]. Each URLLC UE will generate a small payload of 50 bytes according to a Poisson arrival process. Further details of the simulation parameters are provided in Annex 6.2.

In order to balance the load between the macro and the small cell network in the considered HetNet scenario, a cell range extension (CRE) is applied to force more UEs to be connected to the small cell layer. The first step in our evaluation is to identify the optimal CRE value. The results presented in Figure 4-9 show that the optimal user split across the layers is achieved at the CRE value of around 15 dB, which leads to the highest rate of satisfied users which fulfil the latency target. This CRE value will then be used as the input to the DC/MC cell selection criteria.

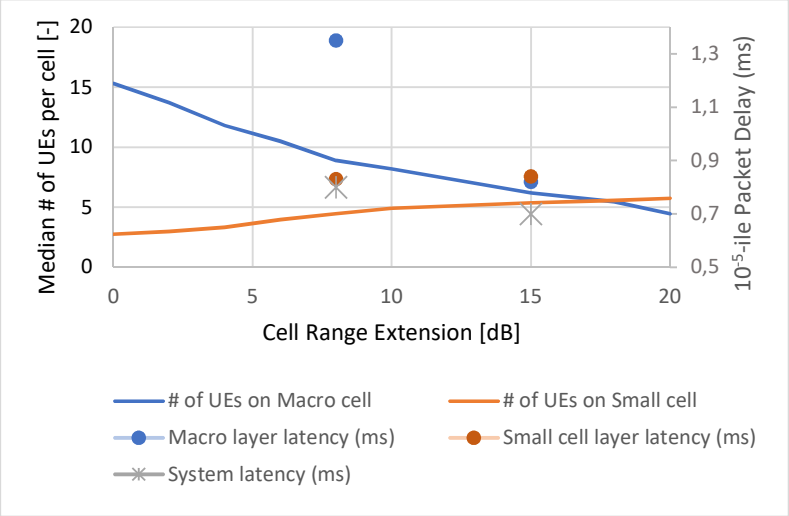


Figure 4-9: Single-connectivity baseline results: Median number of users connected to macro and small cells and packet delay performance at low load for various CRE values.

When enabling dual connectivity, first we carefully select the set of targeted users who should then use data duplication. Specifically, the SCell selection is successful whenever the secondary link for a given UE is relatively strong as compared to the primary link. This is achieved by employing UE assistance in the form of a conventional A3-like measurement reporting event (i.e. "Neighbour becomes offset better than PCell"), according to which the UE compares the level of an SCell relatively to its PCell and reports the triggering SCell including further triggering details. The rationale is that a much weaker secondary link will not only provide any benefit to the targeted user, but in addition will affect negatively other users. Furthermore, we also control the amount of UEs in DC by the parameter *DCRange*, to avoid that the duplication costs exceed its benefits. Thus, the triggering condition for the selection of the secondary cell is computed as follows:

$$\textit{If } SCell_i \textit{ RSRP} \geq \textit{ PCell RSRP - CRE + DCRange then select the } Scell_i \textit{ and enable DC.}$$

For any user for which a SCell was selected (i.e. the UE is configured in DC), its PDCP packets will be duplicated and transmitted via both links. When a PDCP packet is successfully received from a link (i.e. from either the primary or secondary link), we employ a mechanism at the network side to detect the presence of a copy of that packet at lower layers of the other link (e.g. at the HARQ processes) and discard the corresponding PDU to avoid redundant transmissions. It is noteworthy that the discarding procedure introduces cross-layer operations, to maintain the mapping between a PDCP PDU identifier (i.e. sequence number) and the corresponding lower layer PDU(s). This optimization relies on the exchange over X2/Xn of the information of successful delivery status. In addition, we explore the impact as well of increasing the BLER target for the first transmissions (e.g. from 1% to 10%) in connection to data duplication. However, for simplicity we apply the same target irrespective of whether a UE is in DC or not.

Figure 4-10 shows the complementary cumulative distribution function (CCDF) of the one-way URLLC latency for the dual connectivity case (when *DCRange* is set equal to 15 dB) and single connectivity case. The results are provided when the BLER target is set to 1% and 10%. First, it should be noticed that DC provides a latency gain as compared to SC thanks to the data duplication property (i.e. the green lines are better than blue lines). Second, it can be seen that operating at 1% BLER target leads to a lower latency (i.e. solid lines are always better than dashed lines). This is because, due to the very small URLLC packet size, operating at 1% BLER target does *not* require segmentation (implying more transmissions per packet), while on the contrary applying a BLER target of 10% for any UE in the cell, requires a larger number of retransmissions on average to successfully deliver any packet, which in turn increases considerably the experienced latency (above the latency target).

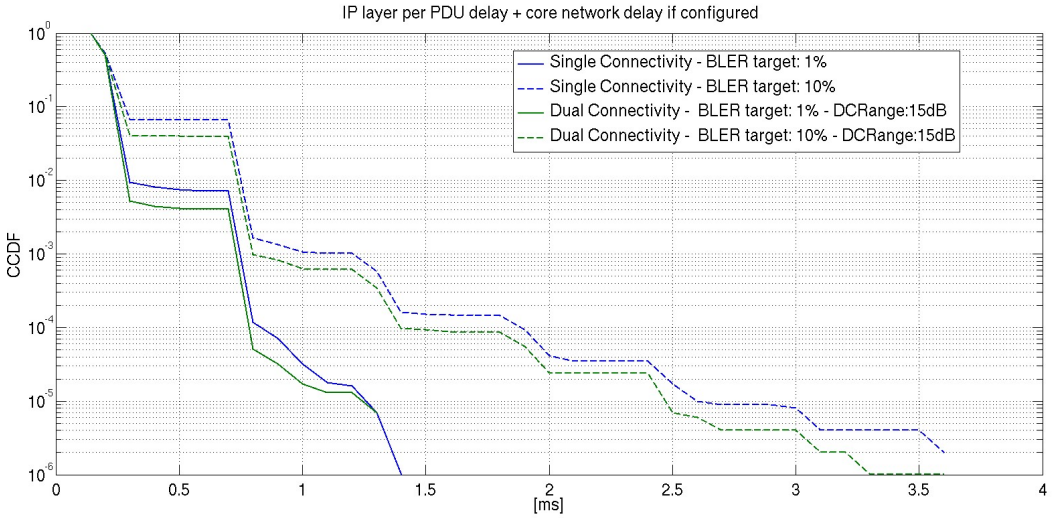


Figure 4-10: CCDF distribution of URLLC latency.

In the future work, we will study further means to reduce the costs associated to data duplication, and therefore means to improve further the data duplication gains.

4.2 Spectrum Management

Before starting to describe the main topics dealing with "spectrum management" in ONE5G, a brief 3GPP summary introducing the usage of new frequency bands, especially in un-licensed bands is carried out. Then, an overview of potential new frequency bands (sub and above 6 GHz) that are candidates for future radio transmission is depicted below. Then, starting from the Shannon capacity formula, it is clear that higher data throughput induces naturally larger signal bandwidth. A way to increase the bandwidth is to aggregate several bands dynamically and/or to use bands in standalone with large spectrum (i.e. millimeter waves for instance). In the case of carrier aggregation, we propose to use spectrum in un-licensed bands that is the main difference with the multi-connectivity approach (see section 4.1).

We split the innovative ONE5G topics about spectrum management in two categories. The first one considers the usage of unlicensed spectrum in standalone mode with 2 topics, including the MulteFire alliance and the second one considering dynamic and flexible carrier aggregation mechanisms using licensed and/or unlicensed spectrum below and above 6 GHz, considering radio resource allocation and network signaling strategies (2 topics are under studies). First results are given showing the interest of using aggregated carriers for increasing the system capacity.

Unlicensed spectrum in 3GPP

The first system that appeared, aggregating LTE licensed and LTE-based unlicensed spectrum was called *LTE-U*, based on 3GPP R10/11/12 but has not been standardized in 3GPP. The unlicensed spectrum was only used for LTE Supplemental Down Link (SDL), typically for offloading. R13 [36.889] has introduced the LAA (Licensed Assisted Access) and LWA (LTE WLAN Aggregation). In LAA the primary carrier uses licensed LTE band whereas the secondary carrier uses the LTE in unlicensed spectrum for supplemental DL only. The LWA uses WiFi in 2.4 and 5 GHz in the unlicensed spectrum for DL only. Then eLAA [RP-162235] (enhanced LAA) was introduced in R14 where unlicensed spectrum is used for SDL and UL aggregation and also eLWA where WiFi in 2.4, 5 and 60 GHz is used for both DL and UL. The MulteFire alliance [MULT] developed a system based on 3GPP R13/R14 that uses LTE-based unlicensed for both DL and UL transmission without any licensed anchor required (in standalone). A study item has been agreed to start in January 2018 [RP-172021] which will focus on New Radio-based access to unlicensed spectrum, possibly including standalone operation with priority to frequency bands above 6 GHz. The unlicensed bands should be used independently for DL and UL and should require no licensed anchor. Then, the system principle of using unlicensed bands with or without licensed ones, has to be determined for future applications. However, the focus of such study item is on providing multi-Gbps data rates rather than securing high reliability and low latency communication.

New 5G frequency bands

Spectrum is scarce and expensive. The World Radio Conference (WRC) has started multiple discussions to identify new frequency bands that should be used in 5G. Especially, WRC 2019 [WRC19] is expected to identify new higher frequency bands (above 6 GHz, typically above 24 GHz) that may be deployed in either licensed or unlicensed networks according to national or regional regulations. Figure 4-11 gives a status about the potential European frequency bands in licensed and un-licensed spectrum that should be used in short, medium and long terms. The 3.5 GHz band is the first one that will be considered for 5Gfor 5G5G NR deployment whereas the 26 GHz band is the main candidate for the micro 5G base station called µgNB.

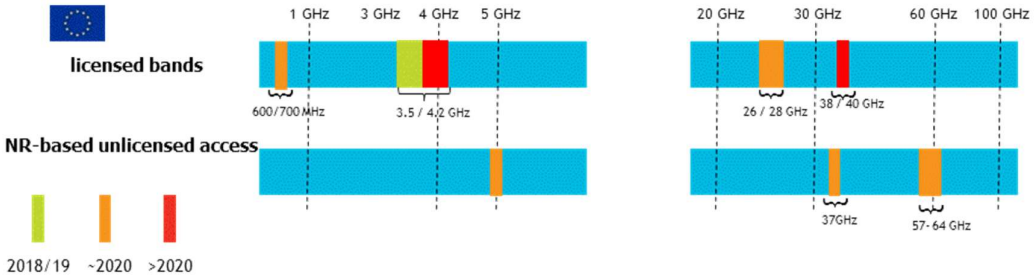


Figure 4-11: Status of the potential European frequency bands in licensed and un-licensed spectrum.

We can notice that the higher the carrier frequency the larger the frequency band and thus the higher the data throughput can be achieved.

4.2.1 Unlicensed standalone operation based on MulteFire evolution

Support for LTE standalone operation in the 5 GHz unlicensed band has been introduced with the first release of MulteFire (MF1.0) [MULTR1]. MulteFire was initially targeted at mobile broadband applications, as it is extensively based on LTE radio specifications. However, MulteFire has recently raised the interest of large-scale technology enterprises in search of wireless radio access technologies that can provide reliable access to private networks globally and without the need for expensive licensed spectrum.

Therefore, the second release of the MulteFire specifications (MF1.1) introduces better support for IoT applications, though enhancements are mainly targeted at massive machine type communication deployments. Additional improvements such as support for autonomous UE mobility and grant-less UL transmissions are also being standardized as part of MF1.1. Nevertheless, enabling highly-reliable and low-latency communication in unlicensed spectrum was not in focus of MF1.1.

An important aspect when considering operation in unlicensed spectrum is that the most restrictive regulatory requirements mandate clear channel assessment based on listen-before-talk. Listen-before-talk 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 for a certain interval of time. If the channel is occupied, the device is not allowed to transmit. Additionally, a device is only allowed to occupy the channel for a limited duration of time before it shall perform a new listen-before-talk procedure. As compared to licensed spectrum, clear channel access based on listen-before-talk obviously represents an additional challenge when targeting high-reliable low-latency communication in unlicensed spectrum.

This is confirmed by our preliminary performance analysis of the reliability and latency using MulteFire. The performance is evaluated in a system level simulator in line with the 3GPP indoor scenario for LAA coexistence evaluations in [36.889]. Since the scope is to evaluate the latency and reliability in a controlled environment, such as a factory, the analysis is for a single operator deployment. A dynamic traffic model corresponding to FTP model 3 in [36.889] is applied. The number of users in one simulation is fixed. FTP data packets of 50 Bytes are generated by each user according to a Poisson process. The network load is adjusted by varying the arrival rate of the Poisson process. Initial simulations only consider 100% downlink traffic. The preliminary results show the impact of load on the distribution of the downlink packet latency. Already for a load corresponding to 1 Mbps of offered data traffic in the simulated network, a reliability of 99.99% can only be achieved for a latency far above 100 ms, as illustrated in Figure 4-12.

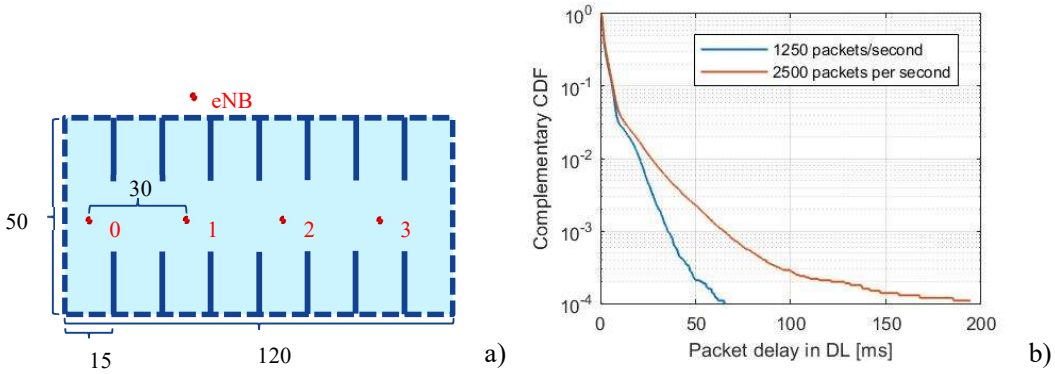


Figure 4-12: (a) Enterprise hotspot scenario. (b) Distribution of the packet delay in Downlink using MulteFire 1.0 and assuming different load conditions.

The target of future research activities is to further assess the latency and reliability performance of other solutions working in the unlicensed spectrum. This includes latency and reliability evaluations in uplink using both scheduled and grant-free transmissions. Evaluations will be performed in a realistic industrial environment with accurate models of traffic and propagation conditions, and will consider traffic mix scenarios with both UL and DL traffic. Specific radio resource management optimizations such as power control, scheduling, and frame adaptation optimizations will also be considered. Then solutions to increase the reliability and reduce the latency of communication in unlicensed spectrum will be studied and evaluated. Example of such solutions may include (but might not be limited to) frequency and time diversity techniques (e.g. duplication and time repetitions), shortened transmission time interval, optimizations for small data transmissions requiring high reliability and low latency, etc.

4.2.2 Unlicensed spectrum management in NR 5G based on KQI

Due to constantly increasing traffic demand, the use of unlicensed spectrum is becoming a key technological asset as well as a challenge for new deployments. Unlike in licensed spectrum, where operators have full control and reserved resources, the design for unlicensed carrier aggregation must face several challenges, including fair coexistence support, multiRAT interference and uncoordinated operation.

As established in LAA, LTE in unlicensed bands can be achieved by adapting the CA feature of 3GPP, where at least one carrier is from licensed band and another is from unlicensed band. Beyond that, in eLAA (enhanced LAA), UL procedures are defined.

In this LAA/eLAA context the channel access technique Listen Before Talk (LBT) with Clear Channel Assessment (CCA) has been defined for coexistence with other RATs (e.g. WiFi) [36.889]. Regarding future standards, 3GPP 5G NR Study is expected for June 2018 [38.889]. Other standards like MulteFire have been created to support standalone unlicensed operation (with foreseeable addition for 5G in future 3GPP releases [R1-1711469]), making also use of LBT/CCA methods. It is furthermore necessary to explore the possibilities of unlicensed channel access schemes proposed by 3GPP to enhance the QoE.

In the framework of this project, KQIs have been defined to better analyze the QoE performance aspects that can be improved in NR. In this way, the KQIs will be used to assess the impact of the indicated channel access techniques [R1-1711465] [R1-1711467] in the QoE of the users. As illustrated in Figure 4-13, this activity has initially the focus on small cell licensed/unlicensed scenario LAA #2 [38.889] (Indoor/outdoor hotspot ONE5G use case), where further 5G NR developments can be added. To this extent, different configuration parameters and procedures selection (for example, UL scheduled or unscheduled) are being simulated to analyze services performance and possible constraints.

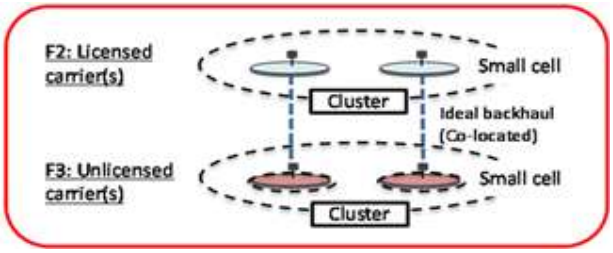


Figure 4-13: LAA Deployment Scenarios and coexistence mechanisms (TR 36.889) Scenario 2.

The impact in the service performance of different situations is addressed, including the coexistence of multiple RATs in the same indoor scenario (e.g. LAA, WiFi) with different load conditions (users/AP/eNB number and distribution and available channels).

Here, the activity will aim to identify suitable configurations for each service in the context of LAA. A sensitivity study of the different channel access parameters will be provided (following the model shown in the Table 4-1), considering different services (eMBB and URLLC 5G[2]) and achievable KQI values.

Table 4-1: Parameters and model of the service-oriented sensitivity and configurations study.

Channel access parameter	Optimal choice according to service type
Energy Dectection (ED) Threshold selection method	Per UE/ Per eNB/uniform
ED Threshold value range	Values to achieve for service KPIs, depending on load, multiRAT interference, context
Contention window size	Depending on scenario
Achievable fairness	Yes/no and how Possible trade-off achievable fairness vs performance
Backoff algorithm	Suggested algorithm
Reservation signal options	RRM/CSI, multiconnectivity, signaling channel
Other parameters	Cross layer (HARQ), RLC and other

Afterwards, specific Machine Learning algorithms will be implemented to dynamically adjust the values of channel access configuration for a given set of KQI requirements. Here, the starting point will be the parameter ranges defined on 3GPP LBT Cat.4 [36.889].

To achieve the objective, it is planned to integrate an unlicensed module to our proposed carrier aggregation manager system (cf §4.1.5) incorporating unlicensed sub 6 GHz bands. Such a module will be integrated in the Component Carrier score calculation, allowing the system to choose between licensed/unlicensed multilink configurations.

These tasks are being developed based on exhaustive simulations of coexistence (multiRAT / multi-operator) scenarios performed in NS3 simulator [PBM11].

As initial step, existing 3GPP standard is evaluated from an end to end perspective with aim to optimize its signaling.

[2] An LTE eNB attempt to access the channel at predefined time instants is called a Tx opportunity. At a transmission opportunity, if transmission has not started, then sensing takes place based on Energy Detection (ED) threshold. If energy is below threshold then the channel is available for transmitting. Otherwise it is considered busy

First introduced in Release 13, Licensed Assisted Access technology allows unlicensed carrier aggregation feature in 5GHz band in downlink shared channel. According to the reports performed [38.889] Cat 4. LBT seems to bring the most promising channel access mechanism to minimize the impact of LTE operating in unlicensed band on other coexisting technologies. By comparison, WiFi standards have been designed from scratch to fully operate on unlicensed band attempting to minimize the impact over other devices. In this regard it is likely that LAA will be intensively used in indoor dense scenarios like the indoor/outdoor hotspot scenario considered in ONE5G.

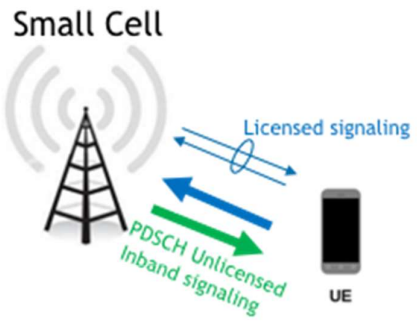


Figure 4-14: LAA signaling scheme.

Given such context, we are considering in this work the LTE in-band signaling contained in downlink shared channel. Such signaling includes cell specific reference signal, synchronization signals, paging and physical broadcast channel. In LAA and eLAA, most of the signaling messages are exchanged using licensed bands. Nevertheless, shared channels as PDSCH do have a specific in-band signaling that cannot be neglected and it would eventually affect the spectrum usage.

On the basis of ON/OFF mechanism LTE Rel. 12, small-cell enhancements were introduced allowing a UE to measure discovery reference signals (DRS) on a deactivated secondary cell (Scell). This procedure enables the network decision process to activate such cells based on UE measurements.

DRS is composed by a set of signals that includes the Primary Synchronization Signal (PSS), the Secondary Synchronization Signal (SSS), the Cell-specific Reference Signal (CRS), and the Channel State Information Reference Signal (if configured). Its duration - although not specified - must be subject to LTE time slot granularity of 1 ms.

As stated by LAA standard procedures, DRS transmission is utilized for cell detection, synchronization, and radio resource management (RRM) measurement within a periodically occurring time window called the DRS measurement timing configuration (DMTC) occasion that has a duration of 6 ms and a configurable period of 40/80/160 ms. However, to reduce a collision probability, the transmission of DRS is also subject to LBT. If data is scheduled during DMTC window, DRS is embedded in data transmission. Otherwise it is sent alone, without data.

Moreover, the modulation and channel coding scheme used in PDSCH is determined from channel estimation obtained from DRS Reference Signals. After measuring RS contained in DRS, UE sends SRS through uplink signaling channel and, according to SINR level a CQI is assigned with a certain Modulation and Coding Scheme (MCS) Rank.

Following 3GPP recommendations in [38.889] we have evaluated the performance end to end in terms of User Perceived Throughput and Mean Delay under different traffic load conditions. Our results probe that periodic sending of DRS signals has a noticeable negative impact on coexisting WiFi (in terms of comparison with the performance achieved when Wifi is coexisting with

another Wifi operator instead of LAA or W+W case) while the performance of LAA stations is degraded as seen in Figure 4-15

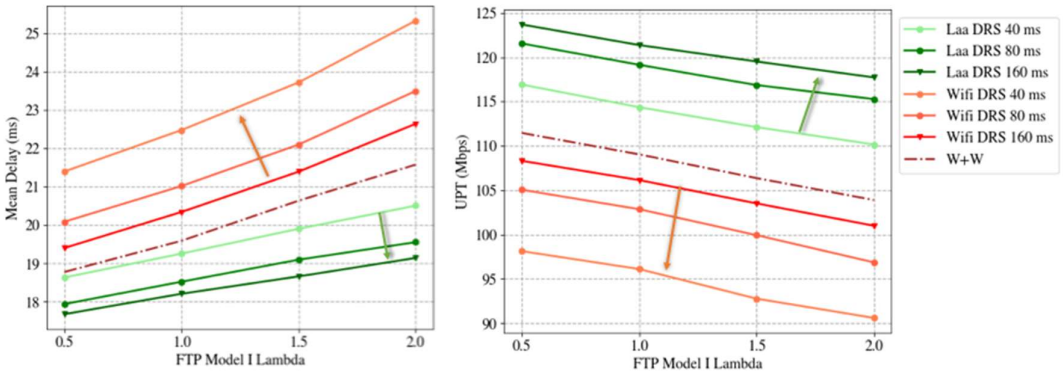


Figure 4-15: Impact of DRS periodicity on WiFi and LAA performance compared to WiFi only scenario.

From these results it can be inferred that enabling DRS signals in a dense indoor scenario can be detrimental for both technologies under high traffic demand conditions.

A first approach to improve performance is to select lower DRS periodicity (160 ms) to achieve improved LAA throughput and delay with less impact on WiFi. But then the comparison may become unfair.

Discussion of other approaches like the temporal suppression of not embedded DRS or dynamic modification of a specific Energy Detection threshold for those signals will be presented as subsequent part of this work.

This protocol design aspect shouldn't be neglected in case of future 5G standards operating in unlicensed spectrum.

4.2.3 Dynamic spectrum aggregation for 5G new radio

In 3GPP new radio (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 [RAN1-91bis], for a UE, a configured DL (or UL) bandwidth part (BWP) may overlap in frequency domain with another configured DL (or UL) BWP in a serving cell. A DL (or UL) BWP is configured to a UE with granularity of starting frequency location of 1 PRB, and granularity of bandwidth size of 1 PRB.

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.

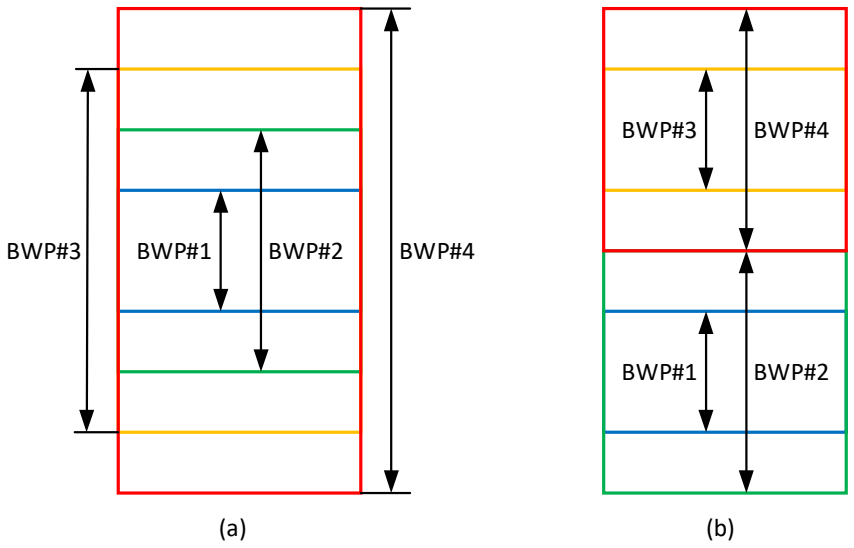


Figure 4-16: Multiple BWP configuration for NR UE, (a) bandwidth adaptation, (b) bandwidth adaptation and load balancing.

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 UE. As shown in Figure 4-16 (a), UE can be configured with four BWPs with same central frequency but different bandwidths, dynamic switching between 4 configured BWPs can adapt the UE receive bandwidth according to the instantaneous traffic throughput demand. Another BWP configuration example is illustrated in Figure 4-16 (b), where 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.

In Rel-15 NR, for a UE, there is at most one active DL BWP and at most one active UL BWP at a given time for a serving cell. In case of paired spectrum, for active BWP switching using at least scheduling downlink control information (DCI), DCI for DL is used for DL active BWP switching and DCI for UL is used for UL active BWP switching. In case of unpaired spectrum, for active BWP switching using at least scheduling DCI, DCI for either DL or UL can be used for active BWP switching from one DL/UL BWP pair to another pair. It has also been proposed to support scheduling DCI with "zero" assignment (i.e. without scheduling downlink or uplink transmission) for active DL/UL BWP switching. And for DL scheduling DCI, UE is expected to send positive HARQ-ACK for zero-size PDSCH transmission.

4.2.4 Optimization of the Radio Resource allocation in traffic/services transmission by spectrum aggregation

The main objective of our study is to benefit from the usage of several frequency bands for increasing the network densification, especially when considering ultra-dense urban environment. As illustrated in Figure 4-17, macro and micro gNB (µgNB) are composing the network deployment allowing to increase the global network capacity. Typically the macro and µgNBs are transmitting at different carrier frequencies, minimizing the interference level. However, µgNBs are using the same frequency implying intra µgNB interference.

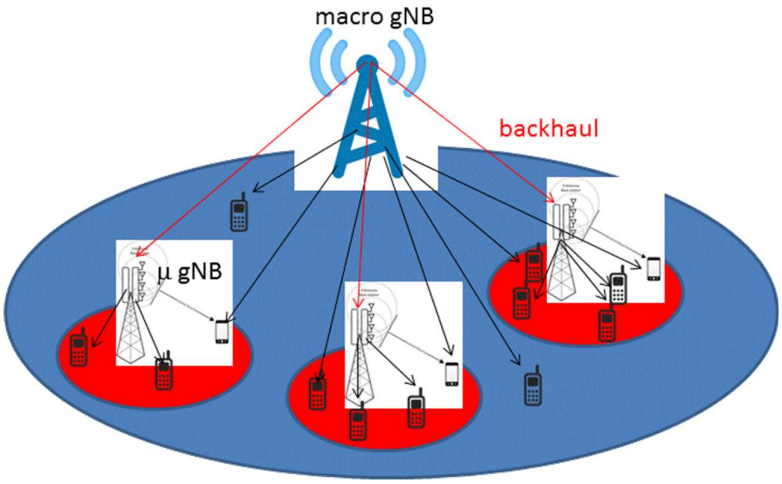


Figure 4-17: Ultra-dense urban environment scenario.

First studies about Heterogeneous Network (HetNet) have been carried out by aggregating 2 different frequency bands below 6 GHz by deploying macro and Femto cells at 2.15 GHz and 3.5 GHz, respectively. The performance evaluation has shown very significant gain in terms of capacity increase as depicted in Figure 4-18.

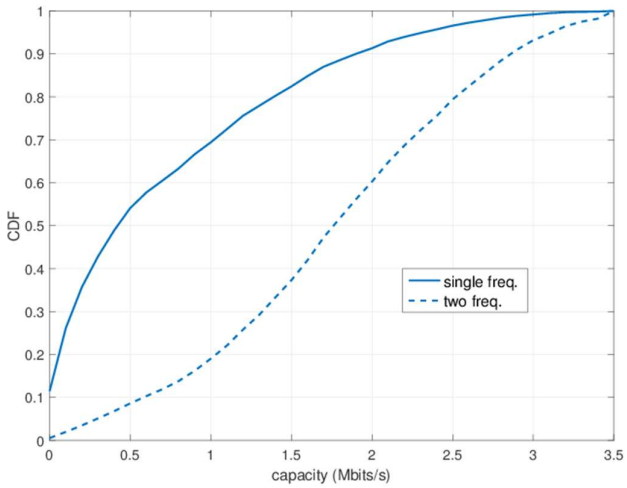


Figure 4-18: CDF of the capacity in a network featuring 108 UEs, 18 femtocells, and 6 UEs per femtocell. Comparison between single and two frequency bands.

Starting from this result, our motivation in the ONE5G project is to use spectrum below 6 GHz for the macro gNB (typically around 4 GHz) and above 6 GHz for the µgNB, using the millimeter wave bands (typically around 30 GHz) to optimize the network 5G KPI in terms of: capacity, spectral efficiency, latency, fairness and energy efficiency. Our study set-up considers that the µgNBs use multiple antennas in transmission (Massive MIMO system studies in ONE5G WP4) for focusing the energy towards the targeted user equipment (UE) minimizing the interference within the µgNB and between µgNBs. Our study amounts to optimize the multi-user radio resource allocation in time, frequency and space (considering Massive MIMO and precoding techniques) firstly within the µgNB using mm-waves and then more largely by using the macro gNB for transmission. The criteria of radio resource allocation will take into account the service type to transmit and also the constraints linked to the service. Indeed, the first research axis is to

optimize the multi-user allocation by spatially grouping the users via the selected set of precoders minimizing the inter-user interference while globally minimizing the number of radio resource blocks (time and frequency allocation) depending on the transmit service type. The dynamic selection or the aggregation of the two bandwidths will be considered in the resource allocation strategies. Each UE can be viewed by either the macro or the μgNB (or both) affecting directly the radio system QoS.

System level performance evaluation with various scheduler types both with full buffer and non-full buffer traffic models according to different criteria depending on the service (capacity, spectral efficiency, latency, fairness, energy efficiency) will be taken into account in the study.

The 3GPP channel model [38.901] for frequencies from 0.5 to 100 GHz has been implemented and validated in our own system level simulator (Matlab based). This channel model has to a large degree been aligned with earlier channel models specified below 6 GHz such as the 3D SCM model or IMT-Advanced. It also conforms to the channel model for frequency spectrum above 6GHz in 3GPP technical report [36.900] and specified several types of antenna array for the Massive MIMO implementation. Our system simulation parameters are aligned with the technical report [38.913], where a total of 12 deployment scenarios have been introduced, including the dense urban one. The spectrum management between the macro and μgNB needs also to define control signaling between them to schedule the UEs in the optimum way. Some hypothesis will be done about the duplex transmission mode (FDD or TDD) operated by the macro and μgNB and their synchronization level.

4.3 Mobility and Load Balancing Optimizations

Recent research in mobility has followed two directions: on the one hand, novel techniques have been proposed, in order to improve the mobility mechanisms themselves, in an attempt to optimize mobility-specific performance indicators, like the data interruption time [GMP16, GMP+17]. On the other hand, mobility management has been studied in order to address certain self-optimization use cases, like load balancing [PRB+11] or traffic steering [PRL+13].

In this project, mobility management will be enhanced following a number of directions: from using context and end-to-end performance metrics as new sources of information, to developing novel user-to-base station association rules and taking advantage of multi-connectivity benefits as an extension of current dual-connectivity.

Mobility status in 3GPP

As of now, 3GPP has addressed mobility optimizations mainly through the DC feature, as a means of providing an always-on radio link for a high-speed moving user [36.842]. In this way, DC has proven to successfully reduce the number of radio link failures and the number of handover failures, especially in scenarios where several small cells overlap the coverage area of a macrocell along a given path. Besides, several enhancements have been proposed in order to reduce the service interruption time, given the impact of this on the TCP performance [36.881], like RACH-less or make-before-break handovers. In NR, it is expected that this time could be reduced to zero [38.913, R2-168299].

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

In this section, a preliminary study related to mobility optimization is described. The global objective of this work is to develop an algorithm to ensure high quality services through a seamless quality of experience traffic steering making use of a predictive network management. This work is focused on eMBB services so that it is specially related to ONE5G use cases no. 5 and 6 [D21]. Specifically, this first study shows how traditional load balancing techniques methods have an impact on the QoE for different services. In addition, a first version of a QoE balancing algorithm is proposed.

This work has been carried out using a dynamic system-level simulator developed in MATLAB. Regarding traffic models, the simulator includes different traffic models: VoIP, Video and Web. The simulator also implements QoE calculations on a user basis [OLI16]. Mean Opinion Score (MOS) scale will be used to quantify quality of experience, ranging from 1 to 5 for the worst and the best experience respectively, as perceived by the user.

In the first study, a hexagonal grid with 57 tri-sectorized cells (i.e., 19 sites) has been used as simulation scenario (see Figure 4-19: (a)). With the aim of an easier analysis, traffic is confined in only three cells (cell 32, 33 and 39). Offered traffic is differentiated by services at every cell, and only one service is offered at every cell at the beginning of the simulation time. More concretely, new users at cell number 32 (C32) only demand Web traffic, while C33 and C39 mainly carry initially Video and VoIP connections, respectively.

As described before, a load balancing method based on handover margin modifications ($Margin_{PBGT}$) is considered. Figure 4-19 (b) shows the load of each cell (PRButil) and QoE for each cell (QoE(C)) when $\mathrm{Margin_{PBGT}}$ (C39,C32) and $\mathrm{Margin_{PBGT}}$ (C39,C33) are swept from 3 to 12 dB (i.e., up to the maximum value). This evolution in $\mathrm{Margin_{PBGT}}$ (i.e., the increase in $\mathrm{Margin_{PBGT}}$ (C39, C32/C33) values) is the expected procedure if a QoE balancing algorithm is applied in this scenario.

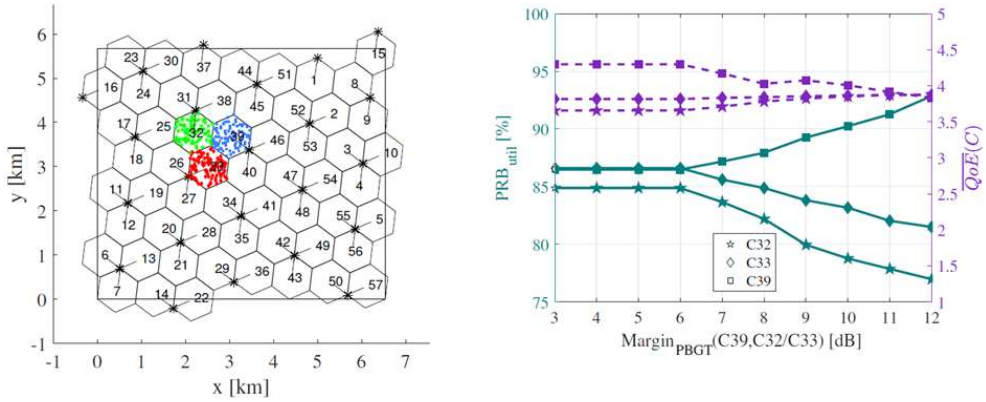


Figure 4-19: Sensitivity analysis of handover margins variations: (a) Simulation scenario and (b) Obtained results.

As handover (HO) margins increases, cell loads are more unbalanced at same time that QoE are more balanced for the three cells under study. This provides a strong indication that load balancing and QoE algorithms cannot be tuned simultaneously. Indeed, a better QoE balance between cells supposes a worse load balance.

In the second experiment, the proposed QoE balancing algorithm is validated. The proposed algorithm is based on tuning handover margins between adjacent cells using a Fuzzy Logic Controller (FLC). These modifications are decided depending on the average cell QoE taken from the previous time period. Eventually, QoE equilibrium will be reached when QoE difference is negligible for every adjacency in the network.

In this experiment, a real LTE network scenario has been developed in the simulation tool (Figure 4-20). Specifically, this scenario is based on a network deployed in South-east Asia. In addition to the eNB positions, the simulated traffic distribution has been taken from this network.

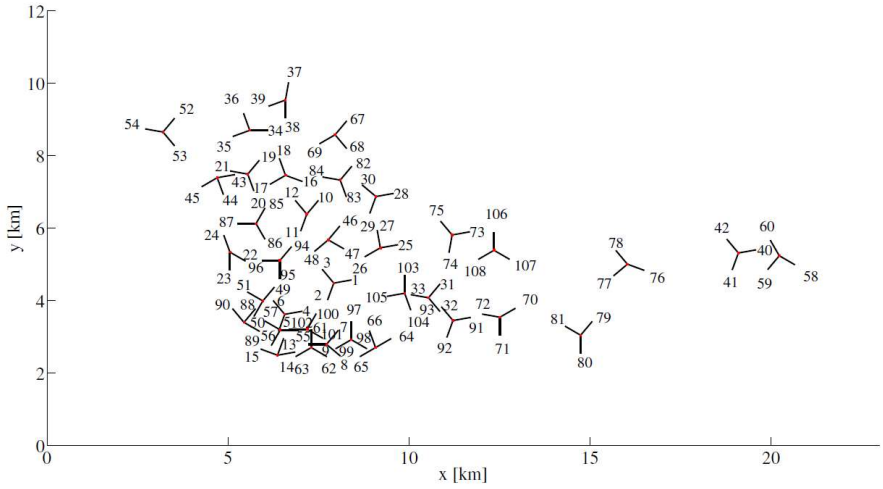


Figure 4-20: Real scenario.

In order to analyze the algorithm performance, a figure of merit has been defined. Specifically, QoE imbalance is defined as the average value of the absolute differences between the QoE of every cell and its neighbors. The QoE for a given cell is computed as the average of the per-service QoEs, which are computed, in turn, averaging the QoE perceived by the users of such service. Figure 4-21 shows the mean value of the absolute value of $\mathrm{QoE}^{\mathrm{IMB}}$ over all the cells in the scenario ($\overline{\mathrm{QoE}}^{\mathrm{IMB}}(\mathrm{C})$). As can be inferred from the blue solid line in the figure, initially, $\overline{\mathrm{QoE}}^{\mathrm{IMB}}(\mathrm{C})$ takes a value of 0.35, which means that the average absolute difference of the quality of experience between neighbour cells is 0.35. After the optimization loops, $\overline{\mathrm{QoE}}^{\mathrm{IMB}}(\mathrm{C})$ reaches 0.18, meaning that the average difference of the quality of experience between neighbour cells has been reduced in 0.17 as it was expected. $\mathrm{Margin}_{\mathrm{PBGT}}$ Imbalance [dB] depicted in the orange dashed line in the figure, shows the mean value of the difference between the value of the handover margin and the default handover margin value (3 dB) per adjacency. At the initial point there is no imbalance, but in every optimization loops, as the $\overline{\mathrm{QoE}}^{\mathrm{IMB}}(\mathrm{C})$ is reduced, $\mathrm{Margin}_{\mathrm{PBGT}}$ Imbalance [dB] increases to maintain the QoE balance. $\mathrm{Margin}_{\mathrm{PBGT}}$ Imbalance [dB] reaches its maximum value of 6 dB when the $\overline{\mathrm{QoE}}^{\mathrm{IMB}}(\mathrm{C})$ is minimum.

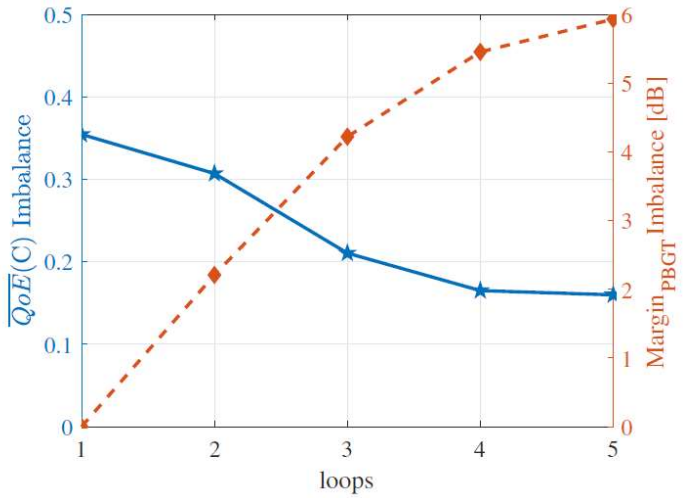


Figure 4-21: Mean Cell QoE and Margin PBGT Imbalance.

In the third experiment, a per-service implementation of the QoE balancing algorithm has been carried out. To reach a finer QoE balanced network in which there are no big differences in the QoE of every service, the handover margin has been divided into four handover margins (one per service). This way, between cell iCiii and cell jCjjj there will be four handover margins: $\text{Margin}_{\text{PBGT VOIP}}(Ci, Cj)$, $\text{Margin}_{\text{PBGT VIDEO}}(Ci, Cj)$, $\text{Margin}_{\text{PBGT FTP}}(Ci, Cj)$ and $\text{Margin}_{\text{PBGT WEBHTTP}}(Ci, Cj)$.

The aim of this division is to modify each one of them with the proposed algorithm to reach a QoE balanced network situation in which QoE per cell and service is also balanced regardless of the service. Hence, there will be no big differences in the QoE of every particular cell or service throughout the network.

With regards to the balancing algorithm, the same controller has been used as in the previous experiment (per-cell QoE balancing) in the same real 108-macrocell scenario, but in a different manner. Here, there are 4 FLCs per cell, i.e., one per cell and service. The input in each of the 4 FLCs is the difference: ΔQoEdiff (i, j, S) = QoE(Cj)− QoE(Ci, S) between neighbour cell Cj and service S of serving cell Ci. There will be 4 outputs, one for each FLC.

Now, cell dominance areas will be reshaped and resized per service for QoE balancing purposes. Within the same cell, dominance areas for each service do not have to be identical.

The figure of merit depicted in Figure 4-22 is:

$$QoE^{IMB}(C,S) = \frac{\sum_{i=1}^{N_C} \overline{QoE}(C_i,S) - \frac{\sum_{\forall j \neq i} \overline{QoE}(C_j)}{N_C - 1}}{N_C}$$

Where N_C is the number of cells considered and $S \in \{\text{VoIP}, \text{Video}, \text{FTP}, \text{HTTP}\}$.

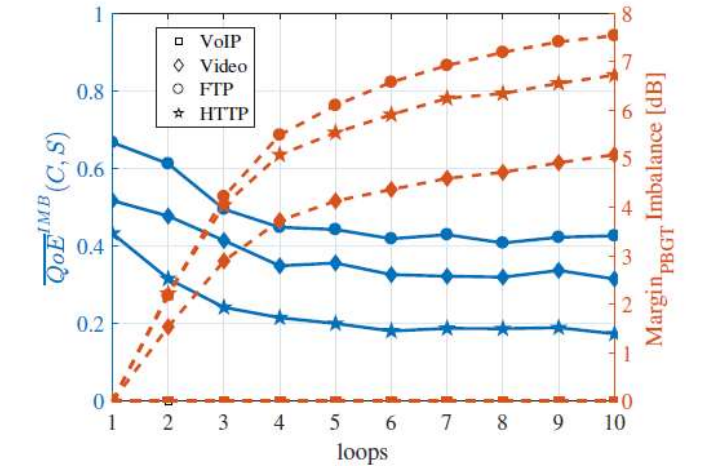


Figure 4-22: Mean Cell –Service QoE and Margin PBGT Imbalance.

In Figure 4-22, there is one blue solid line per service, although there is no information to show that of the VoIP service. VoIP traffic is very little and it is scattered along a few cells in the network, which explains why the algorithm is not capable of making any decision related to this particular service. For the rest of the services, QoE^{IMB}(C, S) is reduced after the optimization process finishes.

Regarding $\text{Margin}_{\text{PBGT}}$ Imbalance [dB], in this particular experiment, there is one orange dashed line per service. That of the VoIP service does not change since the optimization algorithm does not have any information to change handover margins. For Video, FTP and HTTP services,

$\text{Margin}_{\text{PBGT}}$ Imbalance [dB] increases in every optimization loop and reaches its maximum value when $\text{QoE}^{\text{IMB}}(C, S)$ is minimum to maintain the balance of QoE(C, S).

It is worth mentioning that, in collaboration with WP2, the work in this section is proposed for implementation in WP2 system level simulations. Additional results will be then potentially obtained.

4.3.2 Social events information gathering, association and application to prediction in cellular network performance data

In this section, a preliminary demonstration of the gathering and analysis of context information and, particularly, social event data is described [FSB+17] [FDS+18]. Social events (e.g., concerts, sport matches, parades, etc.) are main context variables with an enormous impact in the network performance. Social events typically lead to huge increase in the throughput demand, network elements computational overload and typically, service degradation for the users.

Given the availability of new Internet sources that can provide detailed information about social events (e.g., calendars, social networks, event aggregators) the possibilities of using this data in the management of cellular networks, and especially in its optimization, have hugely increased in recent years [BEC+12].

In this way, a social-aware operations, administration and management (OAM) support system, able to capture both social data information and cellular network information was developed, as shown in Figure 4-23.

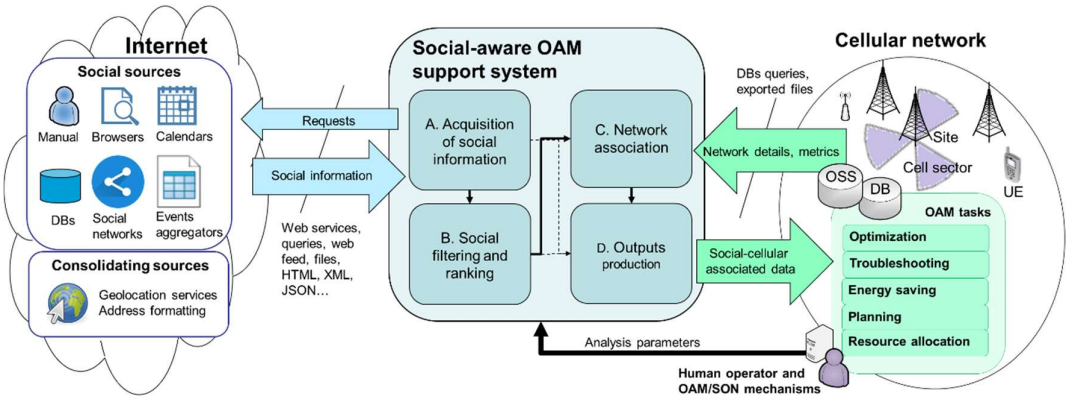


Figure 4-23: Social-aware OAM support system.

These social sources typically provide semantic data of the different events, such as start time, latitude and longitude of their location, type (e.g., musical, parade, sport, etc.), venue (e.g., stadiums, concert halls, etc.), address and popularity. This information acquired from the Internet sources (module A in Figure 4-23), is filtered and ranked based on their expected relevance (module B), and associated/correlated with the cellular network details and data (module C), such as the location and antenna bearing of the base stations, their coverage and the performance impact of previous events. Based on this, the impact of different events can be analyzed/predicted for each base station.

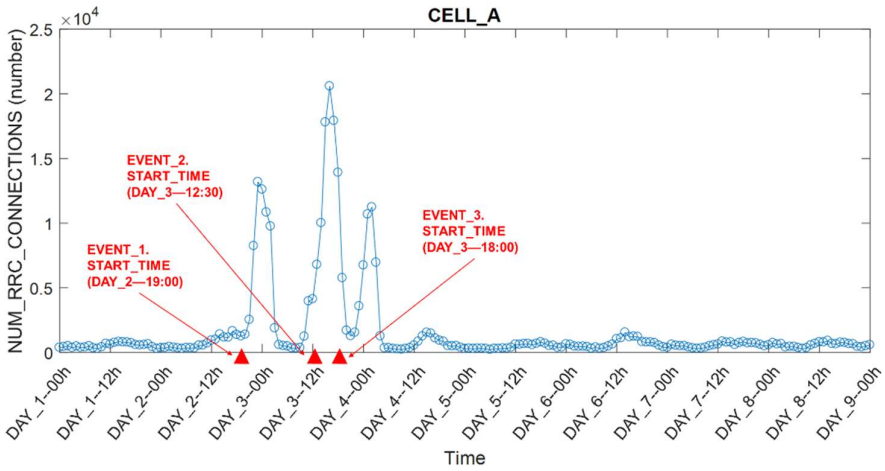


Figure 4-24: Example of the impact of social events in the demand of a cell.

The applications of the generated results (by module D) are envisaged for two main scenarios. Firstly, in classic "slow" OAM (Operations, Administration and Management) activities (e.g., optimization/planning working in periods of hours/days) the resulting information is of key interest to guide pre-emptive resource allocation prior to the events: deployment of temporary base stations, antenna tilt reconfiguration, additional frequency allocation in the area, etc.

Secondly, the applicability of social information in advanced optimization processes, as the ones envisaged in this study, focuses on their use for performance analysis and prediction, as well as on forecasting the throughput demand. In this line, the presence of social events and the analysis of the impact of previous occurrences is key to a proper prediction of the demand and its service characteristics/requirements. An example of this relation is shown in Figure 4-24. Predictions supported by this information can be employed as an input for the configuration of the optimization procedures (especially mobility load balancing).

In the field of the forecasting of performance metrics during future events, a proposed approach is based on the use of nonlinear autoregressive exogenous models (NARX). NARX models imply the forecasting of a *target metric* based on the past values of the target metric y and the current and past values of additional *exogenous input metrics*, x, following the expression $\hat{y}(\mathrm{t}) = \psi\big(\mathrm{y}(\mathrm{t}-1)\ldots\mathrm{y}(\mathrm{t}-\mathrm{n_y}), \mathrm{x}(\mathrm{t}-1),\ldots\mathrm{x}(\mathrm{t}-\mathrm{n_x}), b(t)\big)$, being n_y and n_x the maximum delay of both variables respectively.

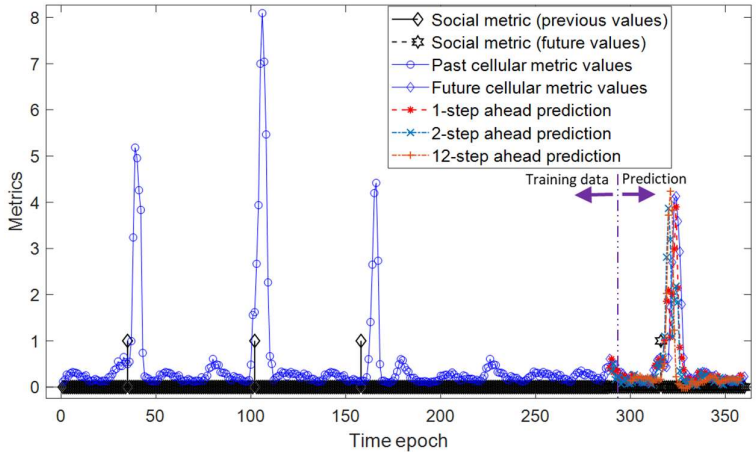
This model fits to our scenario considering the social information, which is typically well known in advance (days ahead), as the exogenous input x. The target metric y can be any performance indicator of the network.

As a baseline approach, the social event information is translated into a binary metric, where a value of '1' indicates the start time of any of the events previously selected as having a possible impact to the performance metric. The rest of the social metric values is set to '0'.

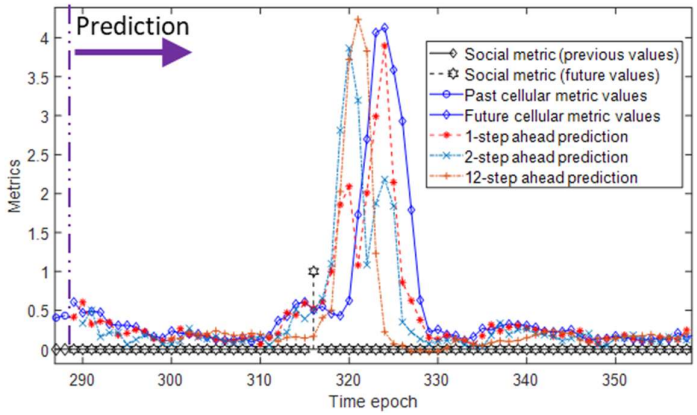
For the currently ongoing study, different configurations of the predictions algorithm are being tested based on real LTE cellular network metrics (processed to increase the number of available cases), showing good results in the application of shallow neural networks for long and short time forecasting. Figure 4-25 shows an example of this, where the approach is applied for the prediction of a cellular metric at different number of *steps ahead*, this means, for a different number of previous unknown values of the metric. This is a typical way to evaluate the quality of the prediction of values beyond the known data. In this way, 1-step ahead indicates that the

estimation of one value of the metric ($\hat{y}(t)$) for a certain instant (t) has been tested knowing its immediate previous value ($y(t-1)$) and more before it ($y(t-1) \dots y(t-n_y)$). 2-step-ahead prediction indicates that for the prediction $\hat{y}(t)$ the most recent known value (measured, not estimated) is $y(t-2)$.

Figure 4-25 shows how the applied method provides very similar profile and values of the expected metric in respect to the measured one, even for 12-step ahead prediction. It can be also observed how this prediction has around 1-hour time-shift in respect to the real values, given the difficulty to perfectly predict the instant where the demand will start to grow so many samples before the event itself: the growing demand curve might began a little sooner or little later in respect to the start time (represented by the "social metric" in the figure) depending on the specific event. Long-term optimization mechanisms based on these predictions can take into account the general moment of occurrence and the expected metric profile/values. Closer to the event, their parameters can be fine-tuned via short-term predictions.



a) Training past data and predicted data in comparison with future values of a cellular metric.



b) Detail on the predicted values in comparison with real values of a cellular metric.

Figure 4-25: Example of different number of steps ahead predictions of a performance metric showing both the training and the predicted data as well as the social metric used as exogenous input.

Further developments will assess the performance of the approach and it might consider possible optimizations in the definition of the parameters of the neural network, as well as the use of multivariate models in terms of social and cellular input metrics.

4.3.3 Multi-access Edge Computing (MEC)-aware UE-cell connectivity

5G networks will need to effectively support huge amounts of traffic streams, both heterogeneous in terms of performance requirements, as well as variable in space and in time. In addition, the emergence of MEC will introduce computing capabilities at the edge of the network and will provide an open environment targeting low packet delays due to close proximity to end users [EFS18]. Moreover, the everyday expanding applications deployed on various platforms, lead to the presence of complex processing algorithms to be conducted (e.g. video analytics, augmented reality). Consequently, task offloading can help to relieve user devices from complex computation via offloading those tasks to a nearby MEC server. From that perspective, the applied rule for user-cell association plays a key role towards efficiently exploiting the entire set of resources (i.e. radio and computational). Nevertheless, current mobile systems have been planned and deployed so far by following traditional paradigms of network planning (e.g., based on radio-only coverage). Unfortunately, this approach is not sustainable anymore, as current cell association rules completely discard the aforementioned availability of processing resources at the network's edge, hence, they fail to constitute cost-effective and flexible solutions for efficient network deployment and operations.

Inspired by the above described situation, in our work [EFS17], focusing on a K-tier cellular network, assuming that the BSs and UEs are placed in random locations, and BSs belonging to different tiers are assumed to have different transmit power capabilities, spatial densities (i.e., BS/unit area) and available processing powers we propose a new, MEC-aware cell association metric that aims at minimizing the execution time of an offloaded task at the MEC server, along with ensuring connectivity to the closest BS. This is motivated through questioning the optimality of the conventional, maximum DL RSRP-based association rule, when it comes to the latency needed to complete an offloaded task. The performance metric of focus, termed as the Extended-Packet Delay Budget (E-PDB) is a one way delay between a user terminal wishing to offload a task and a MEC host collocated with the connected radio node, thus, it is composed of two delay components, namely, the radio transmission latency (in the UL) along with the computational latency, i.e., the time needed to accomplish the processing task at the MEC host. As a first step, we consider an inter-tier equal allocation of the radio and computational resources, dependent on the average number of connected users to a given tier.

With the aim of visualizing the impact of the two studied association rules (i.e., the existent, 3GPP-based one and the proposed MEC-aware one which takes into account the processing capacity of the MEC servers), from a coverage area standpoint, a two-tier network deployment can be observed in Figure 4-26, where the solid blue lines represent the radio coverage regions determined by means of applying the maximum DL RSRP association rule, while the dashed red lines are those determined by means of applying the proposed, MEC-aware UE-cell association rule. The dashed lines represent the association of a given UE to its serving BS. In this figure, the maximum transmit power disparity between a tier-1 BS and a tier-2 BS is 8 times larger than the computational disparity between the two tiers, i.e., the dissimilarity of the MEC servers collocated with a macro/ micro BS. This parameter is referred to as the *radio/ MEC disparity ratio* and in what follows it is termed as parameter ω. The grey shaded users are the ones for which execution of the two cell association rules results to different BS/ MEC nodes for connectivity.

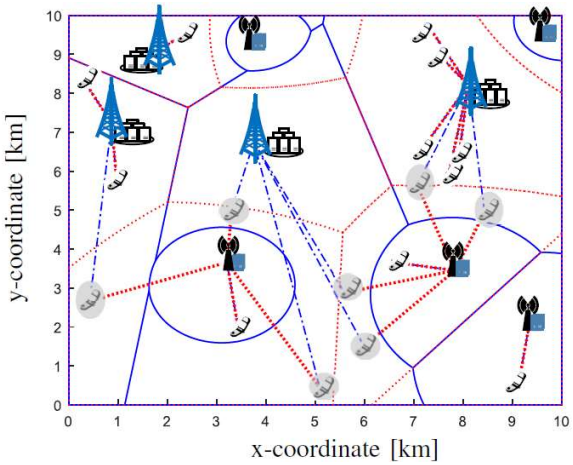


Figure 4-26: A zoomed realization of a two-tier network consisting of macro and micro BS.

To highlight the latency reduction realized via the proposed association metric, the Complementary Cumulative Distribution Functions (CCDF) of the experienced E-PDB by applying the two metrics for 30 users per unit area with randomly generated tasks, along with the percentage of decoupled association decisions for the UEs are depicted in Figure 4-27 (left). For more information regarding the system model and the simulation environment, the reader is kindly referred to [EFS17]. We observe that for values of ω greater than one, the proposed computational proximity- based association rule (denoted as "MEC") provides a lower probability to violate a given E-PDB threshold with nearly 60% E-PDB reduction for the 50^{th} -percentile of UEs. This occurs due to the enhanced balance between the proximity and available computational resources at the MEC node. On the other hand, as ω takes values lower than one, the performance is flipped, as the RSRP rule provides a lower experienced E-PDB with the same gain. Consequently, we observe that with the proposed metric, an adaptive, radio/ MEC disparity-dependent association procedure should be considered to minimize the experienced E-PDB when offloading a demanding task to a MEC host. To achieve that, the UE is only ought to acquire knowledge of the radio and MEC disparities. For $\omega = 1$, similar E-PDB performance is achieved when applying one or the other association rule, due to the full overlap of the coverage areas drawn when applying one or the other association rule.

Additionally, the percentage of UEs for which the maximum DL RSRP and the proposed MEC-aware cell association rules provide different connectivity recommendations, is illustrated, as a function of the value of parameter ω in Figure 4-27 (right). As anticipated, for the increase of cross-tier disparity between the radio and MEC capabilities (i.e. $\omega \neq 1$), the two coverage areas become highly divergent, thus, leading to a higher probability of a UE being present in this disjoint region (e.g., nearly 40 % of UEs will reach different decisions upon associating to an BS/ MEC node for large disparities of ω = 0:01 or ω = 80). On the contrary, for the ω = 1 case, the radio and MEC coverage areas will be identical, hence, the application of the two investigated association rules will provide the same preference for UL connectivity.

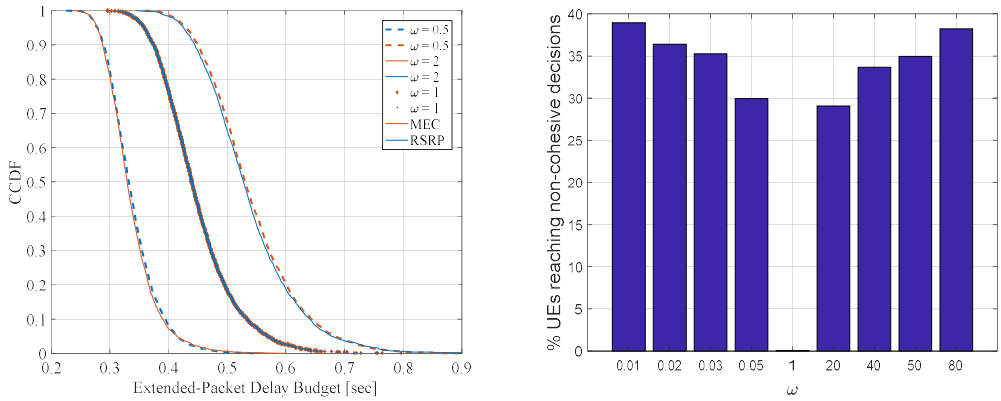


Figure 4-27: (Left) E-PDB CCDF for the two investigated association metrics of different radio and MEC disparity values. (Right) Fraction of UEs reaching non-cohesive decisions upon cell association, as a function of cross-tier radio and MEC disparity.

With the aim of observing the effect of deployment density on the experienced E-PDB, Figure 4-28 depicts the probability of violating a target E-PDB of 0.4 seconds for an increasing ratio of micro-over-macro BS spatial densities ($\frac{\lambda_2}{\lambda_1}$) when $\omega = 2$. We observe a nearly constant association-based outage reduction in favor of the proposed MEC-aware association rule, similar to the latency reduction observed in Figure 4-27. The decreasing slope of the two curves is expected as the number of micro BSs over a unit area increases. This is due to the increasing probability for a UE to be associated with a closer node, thus leading to lower E-PDB values.

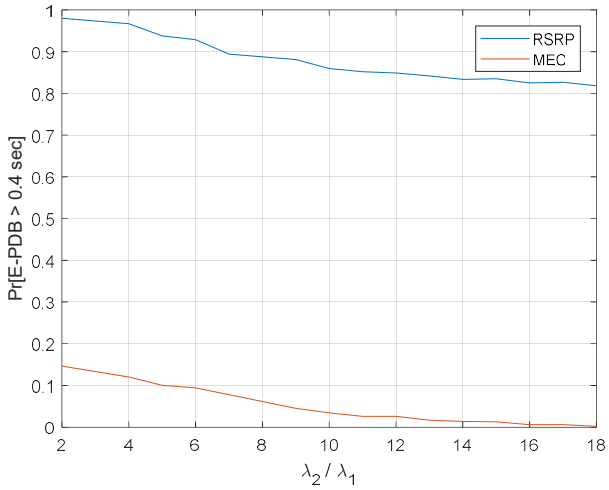


Figure 4-28: Probability to violate a target E-PDB (0.4 sec) as a function of the ratio of micro to macro deployment densities.

4.4 Performance Optimization for UEs with D2D Schemes

This section investigates the benefits of D2D communications from different perspectives. With the increase of density of users in the network as well as the proliferation of bandwidth demanding applications, the networks will be saturated. In this context, D2D communications can be seen as an efficient technique to offload the traffic, increase the coverage and reduce the interference. In addition, D2D communication is of particular interest to provide access and connectivity to the users and devices in the context of Underserved Areas [D21]. The work described in this section

is then beneficial for ONE5G use cases no. 3, 4, 5 and 6 defined in WP2 [D21]. In this section, three aspects related to D2D communications are investigated. The first problem concerns the interference management in a D2D network where the fundamental performance limit, in terms of queuing stability region, is characterized. This has the benefit to find the maximum eMBB traffic load that can be supported by a D2D network. The main motivation of the aforementioned study is to provide a comprehensive performance analysis of D2D networks, which will help to understand how D2D can improve the performance of cellular networks. The other two aspects studied in this section consist of using D2D communication to relay the signals between the BS and the devices respectively for eMBB and mMTC applications. In the case of eMBB services, an opportunistic relay selection strategy that improves the network performance (queuing stability region) is briefly mentioned and will be discussed in more detail in the next deliverable. In the case of mMTC applications, the goal is to use D2D relaying in order to extend the coverage, minimize the power consumption of the UEs and allocate efficiently the radio resources. For that, different topologies and configurations of relays and MTDs (Machine Type Device) are investigated taking into account the connection density and the traffic volume.

D2D status in 3GPP

The standardization work on D2D technologies in 3GPP has started few years ago and was mainly focused on a set of use cases that are of interest for both public safety and commercial mobile networks. For example, proximity services (ProSe) use cases are specified in 3GPP [22.803] and the corresponding architecture enhancements are investigated in 3GPP [23.703]. The feasibility of providing D2D ProSe via LTE has also been evaluated in 3GPP [36.843]. Several D2D aspects related to device discovery, device communication and device relaying have been considered or are under consideration in 3GPP. For example, there are two direct proximity discovery protocols being currently specified in 3GPP [23.303]. Furthermore, D2D relaying has been considered in Release 13 by allowing a UE to operate as a relay (at layer 3) for another UE. Motivated by the development of IoT and wearables applications, the aforementioned relaying scenario will be further enhanced in 5G. In this regard, a new release 15 study item “Study on Further Enhancements to LTE Device to Device, UE to Network Relays for IoT and Wearables” has been opened recently in 3GPP [RP-170295], with the main objective that a UE can act as a relay for remote UEs and mMTC devices.

4.4.1 Performance Analysis of D2D Networks with Interference Alignment

Enhancing the capacity of the network is one of the main promised gains of D2D technology. In order to understand the benefits of D2D in wireless networks, we provide in this section a fundamental performance analysis of a network composed of D2D pairs, using eMBB services, where each device is equiped with multiple antennas. The scenario considered in this section is related to use cases 4, 5 and 6 defined in WP2 [D21]. Since the traffic is usually bursty, the performance is characterized in terms of queueing stability, which a common metric used to find the performance limit of a network in this case. Since the D2D pairs use the same frequency band, managing the interference in the network is of paramount importance. Interference Alignment (IA) and singular value decomposition (SVD) MIMO precoding are two techniques widely used in the literature to deal with interference. In fact, IA is an efficient linear precoding technique that often achieves the full “degrees of freedom (DoF)” supported by the MIMO interference channel. Each technique requires different signaling overhead (mainly related to the Channel State Information (CSI) knowledge at the transmitters). However, it is not clear which solution is more efficient when the signaling overhead is taken into account. In this work, we provide a rigorous comparison between these two techniques taking into acount the channel probing/feedback cost and the dynamic traffic pattern. The interest of this work is therefore to provide a performance analysis of D2D networks and to find the best scheme to use among IA and SVD.

We consider a time-division duplex (TDD) system where the CSI is acquired by the transmitters by decoding the pilot signals sent by the receivers. For SVD precoding, this information is

sufficient to perform the precoding whereas for IA a global CSI knowledge is required. In this case, the transmitters have to quantize their estimated CSIs and exchange them over a wireless backhaul link with limited capacity. This work studies hence the impact of the limited feedback on the system performance taking into account the time varying nature of the traffic. In particular, we provide a rigorous characterization of the queuing stability region of IA and SVD under the limited backhaul constraint and provide conditions under which IA outperforms SVD. In addition, we investigate the performance loss, due to the limited backhaul, with respect to the ideal system with unlimited backhaul. In this section, we provide a brief description of the system model and the main contributions of our work. One can refer to Appendix 6.4 and [DAD+18] for more details.

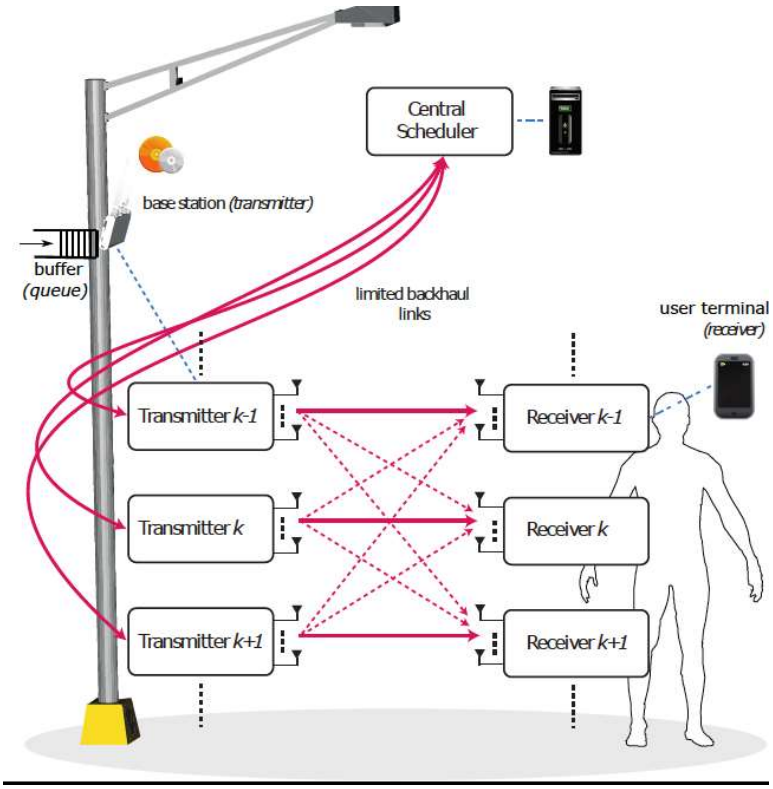


Figure 4-29: Example of the system model

System Model and Performance Analysis

We consider the MIMO interference channel with N transmitter-receiver pairs shown in the figure above. We assume that all transmitters are equipped with multiple antennas. A subset *L(t)* of pairs is active at *time-slot t*. The set of active pairs is decided using a scheduling strategy that should be designed. Each transmitter communicates with its intended receiver and also creates interference to other unintended receivers. When the number of active pairs is higher than 2, IA is applied. However, if the scheduling decides to activate only one pair at each time, the resulting system is a TDMA system where the transmission between the transmitter and the receiver is a simple point to point MIMO system. We assume a TDD transmission strategy, which enables the transmitters to estimate their channels toward different users by exploiting the reciprocity of the wireless channel. As alluded earlier, global CSI is required at each transmitting node in order to design the IA. Each transmitter k has then to quantize its locally estimated channels and to exchange them among each other over the backhaul. The quantization is done using a codebook

with a number of quantization bits equal to B. This quantization has certainly an impact on the accuracy of the assigned rate to each transmitter k. As mentioned earlier, in this work the channel acquisition cost is taken into account. Since we consider a system under TDD mode, the local CSIs are acquired by each transmitter by decoding the pilots sent by the users. This assumes that the users use orthogonal sequences with lengths obviously proportional to the number of active users. Therefore, the amount of time resources required to send the pilots to $L(t)$ active pairs is proportional to $L(t)$. In each time-slot, a portion proportional to $L(t)$ is reserved to pilots and the remaining time is used for data transmission.

Regarding the traffic model, for each user, we assume that the incoming data is stored in a respective queue (i.e. buffer) until transmission. The vector of number of bits arriving in the buffers is an i.i.d. in time process and independent across users. At each time-slot, the set of scheduled pairs should be then based on the queue lengths and the wireless channels of the users. In order to detemine the best performance of the system, one should find the best scheduling policy for both schemes (IA and TDMA). The performance metric considered here is the queuing stability region. The stability region is defined as the set of vectors of mean arrival rates for which the queues of the users are strongly stable. Furthermore, a scheduling policy that achieves this region is called throughput optimal.

We have obtained a precise characterization of the stability region of the aforementioned network model when IA and TDMA-SVD are employed taking into account the limited capacity of the backhaul. We have found analytically that, by taking into account the queue lengths and the CSI acquisition overhead, there is an optimum number of pairs L_I that should be scheduled in the IA context and have provided an exact characterization of L_I. By increasing the number of active pairs L beyond L_I, the set of achievable average rates will not increase. Furthermore, we have also found the best scheduling policy that achieves the stability region of the network. This policy is based on the max weight rule. Finally, we have conducted a comparison between IA and TDMA systems and have provided conditions under which IA yields a queueing stability gain with respect to SVD. In appendix 6.4, a description of these results is provided. More details on the proof and interpretation of the above results can be found in [DAD+18]. In Figure 4-30, we provide simulation results that compare between IA and TDMA-SVD for different numbers of CSI quantization bits B. The results show clearly that IA outperforms SVD when a high backhaul capacity is available, whereas for small capacity (which results in a small number of quantization bits B) it is better to use TDMA-SVD.

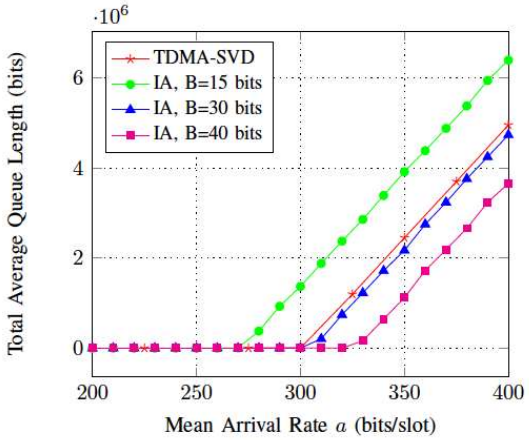


Figure 4-30 Total queue length vs mean arrival rate a; N= 6 transmitter-receiver pairs.

4.4.2 D2D Relaying: Traffic-aware Scheduling and feedback allocation

After analyzing the performance of D2D networks in Section 4.4.1, we consider in this section the scenario where the D2D communication serves as a cooperative relaying scheme in order to help the downlink communication in a wireless network. The focus here is on eMBB applications. In this context, selecting the device that can serve as relay is of paramount importance. This relay/device selection is called scheduling and must depend on the traffic pattern and radio conditions. Specifically, we consider the scenario where 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. We propose a decentralized feedback allocation and relay scheduling algorithm, made at the user side, that takes advantage of their local CSI knowledge in order to achieve higher gains. The performance of this algorithm, in terms of queuing stability, will be compared 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. This work will be described in the next deliverable. One can refer to [DAD+18b] for more details.

4.4.3 Minimizing power consumption and extending coverage using D2D schemes for mMTC services

Massive Machine Type Communications (mMTC) applications are characterized by a large number of devices, low mobility, small and infrequent data transmission, long battery life, low device cost, and low complexity.
Using cellular networks for mMTC applications can lead to network congestion and increased cost of deep indoor coverage. D2D is a promising technology that aims to connect two or more devices directly without going through a direct link between the MTC device (MTD) and the Base Station (BS). D2D relaying can reduce the power consumption, extend the network coverage and have a more efficient use of radio resources [AWM+14].
Our main objective is to investigate different topologies in terms of relays and MTDs and search for an optimal configuration based on the connection density and the traffic volume. The goal is a joint optimization: coverage extension and radio resource efficiency, while minimizing UE power consumption in mMTC scenario (with highly energy constrained IoT devices). D2D technology might be particularly relevant for use case 3, “Non time-critical processes and logistics (factories and smart cities)”, and use case 4, “Long range connectivity in remote areas with smart farming application”, which applies specifically to Underserved Areas scenario where one of the key challenges is the coverage extension [D21].

Three different coverage scenarios can be considered (see Figure 4-31):

- Scenario 1: the Relay and the MTD are in-coverage. In this scenario the relay can help to reduce the power consumption.
- Scenario 2: the MTD has only down-link coverage
- Scenario 3: the MTD is out-of-coverage

In scenario 2 the relay can help to increase the uplink coverage while in scenario 3 the relay will help to increase both down-link and uplink coverage as well as reduce the MTD energy consumption.

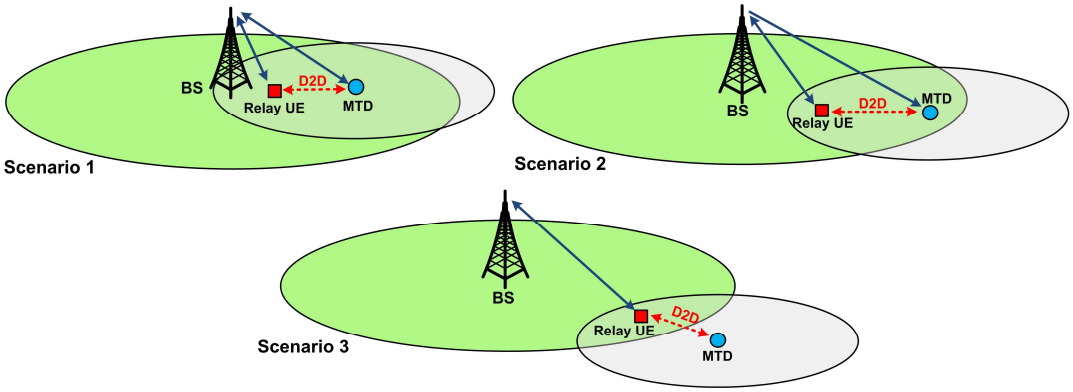


Figure 4-31: Considered coverage scenarios

In a D2D relaying scheme we can distinguish three different phases (see Figure 4-32:):

- The synchronization phase: for scenarios 1 and 2 the MTD can synchronize directly with the base station, using the legacy system. For scenario 3 we will have to specify how the MTD can synchronize with the User Equipment (UE).
- The device discovery protocol: there are two direct proximity discovery protocols being currently specified in 3GPP [23.303]: Model A ("*I am here*") and Model B ("*Who is there?*").
- The D2D communication phase

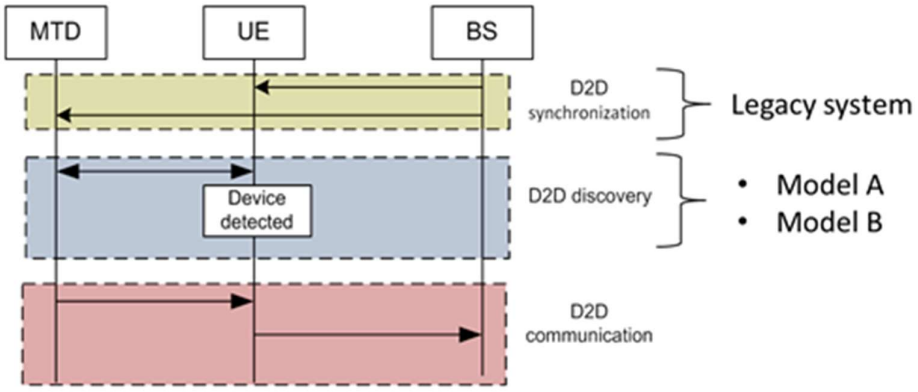


Figure 4-32: D2D scheme

In a first study, considering that a UE can act as a relay, helping in the connection of MTDs with the base station, we proposed a simplified energy consumption model for coverage scenario 1. This model is inspired by the models proposed in [MAH16] and [45.820], but focus only on the communication phase.
One of our main objectives is to identify criteria for establishing the interest of using D2D relay mechanisms, looking for the optimal configuration minimizing the energy consumption.
In our preliminary study, we compare the energy consumption in direct and relaying mode expressing it as a function of the distances between the BS and the MTD and between the BS and the relay. We consider a coverage scenario where both relays and MTDs are in-coverage. In this case, a MTD can connect to the network in two ways: directly to the BS or using a relay. In order to build a simplified model the following assumptions are made:

- Cellular infrastructure supports the discovery and synchronization phases (Network assisted approach). Thus, all devices are synchronized with the BS. We analyze only the energy consumption in data transmission phase.
- MTDs receive the data directly from the BS, i.e. downlink without using a relay.
- Relays are devices that could help MTDs to connect to the network, for which they receive and retransmit the packets coming from MTDs. Our analysis is focused on uplink since mMTC applications mainly generate traffic in uplink.
- Relays are connected to the network using cellular links and MTDs are connected to the relays using D2D links.
- D2D links and cellular links are orthogonal, there is no interference between these systems.
- The communication channel is well known by devices. That means a perfect dynamic adaptation of the modulation and coding scheme.
- Devices transmit at a fixed power considering no power control.

We use a simple model for the calculation of the energy consumption, following the methodology described in [45.820] for the MTD and aligning on the energy consumption model described in section 2 for the UE.

An MTD can operate in one of three modes:

- **Active Mode:** In this mode, the device is able to perform both Rx and Tx activities, which have different power requirements. Radio resources may be assigned for downlink or uplink data transfer.
- **Idle Mode:** In this mode, the device maintains the accurate timing by keeping RF frequency reference active and no radio resources assigned to the device for the downlink or uplink.
- **Power Saving Mode:** In this mode, the device releases all radio connections, only the sleep clock is expected to be running. It is unreachable from the base station. The device is programmed to wake-up periodically to read paging channel.

Our first results (introduced in Annex E) have shown the interest of a D2D relay mechanism compared to a direct link with an unfavorable link budget (MTD far from the BS).

We are currently working on an enhanced energy consumption model in a Poisson Point Process (PPP) scenario. We are still focusing on the transmission phase, considering that the MTD and UE acquire synchronization from the base station but the simplified energy consumption model will be enhanced considering the following scenario and hypotheses (the considered network model is illustrated in Figure 4-33):

- The MTD transmits its packets to the UE and gets a reply with an ACK if they are successfully received. Then, the UE relay MTD data to the BS through a cellular link.
- An Automatic repeat request (ARQ) scheme is used in the MTD-UE link. An MTD will retransmit a packet until it will be successfully received by the UE. We compare the performances of conventional ARQ scheme and Chase Combining HARQ (CC-HARQ) scheme.
- D2D links and cellular links are orthogonal, there is no interference between these systems but interference is possible in MTD-UE links (reuse of resources).
- Different Modulation and Coding Schemes (MCS) are considered for the MTD.
- The locations of MTDs and UEs form two independent stationary Poisson point processes (PPPs).
- For the channel model, a Rayleigh block-fading channel is considered. In others words, the channel remains constant over a single packet transmission but changes independently from one transmission to another. In a next step shadowing will be considered as well.

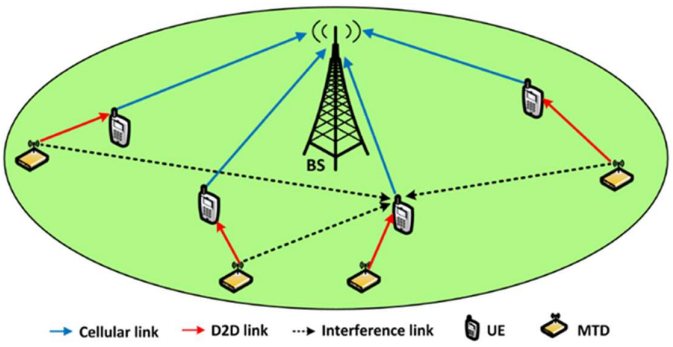


Figure 4-33: Network model

As illustrated in Figure 4-34 the average MTD energy consumption can be calculated as the MTD energy consumption per retransmission multiplied by the average number of retransmissions. The energy consumption per retransmission can be seen as the sum of the energy consumption during the transmission phase, the energy consumption during the reception phase and the energy consumption during the random retransmission delay where the MTD is in idle state: $E_{m,1} = E_{m,T} + E_{m,R} + E_{m,I}$. The random retransmission delay between two transmissions allows the model to satisfy to the independency hypothesis between two retransmissions.

The average global MTD energy consumption can then be written as:

$$\overline{E_{m,global}} = E_{m,1}\overline{\mathrm{T}}$$

where $\overline{\mathrm{T}}$ is the average number of retransmissions. $\overline{\mathrm{T}}$ depends on two variables, the MTD density (devices / km^2), and the MTD SINR. .

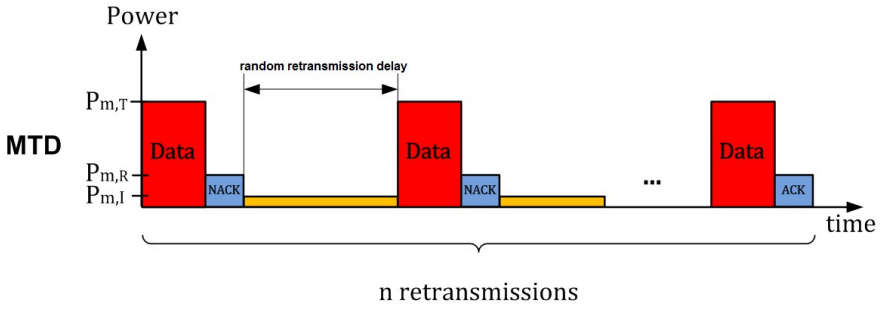


Figure 4-34: Energy consumption modelling considering ARQ and CC-HARQ schemes in a PPP scenario

An important future work is to extend the energy consumption model considering the device discovery phase. Other coverage scenarios (where the MTD can be out of coverage or at the coverage edge) will also be studied, investigating a new synchronization process where the UE should transmit the synchronization signal.

5 Concluding Remarks

In Chapter 2 we outlined our assumed architecture and protocol stack reference models that will be used throughout the ONE5G project. The RAN user plane protocol stack was first described, including the PHY, MAC, RLC, PDCP, and SDAP layers in line with the 3GPP NR agreements. The related QoS architecture was presented, describing how E2E data flows are mapped to QoS flows and DRBs, including related responsibilities and types of QoS information available at the different protocol layers. Special emphasis was put on describing the new functionalities as compared to the LTE protocol stack and QoS architecture. As an example, it was described how the combined QoE manager (aka application layer scheduler) in the SDAP layer and the faster radio-aware packet scheduler in the MAC-layer plays an important role in controlling the E2E performance and hence improve the KQIs for different services. The mapping and scheduler related functionalities at the SDAP and MAC layers must therefore be carefully designed in harmony to achieve the best possible E2E performance. The concept of the new 5QI table was also outlined, i.e. the new QoS parameter table that comes with the 3GPP NR.

Moreover, different D- and C-RAN network architectures were described, including C-RAN architectures with either high-layer or lower-layer functional splits. For the lower-layer split options, both options with CPRI, eCPRI, and NGFI were outlined in line with recent NGMN publications. Those C-RAN options will be adopted as the ONE5G reference models. Also, the concept of E2E network slicing was introduced.

Further, the new three-state RRC machinery was outlined. Describing both the many new attributes of this control plane protocol as current designed by 3GPP for the NR, as well as identification of the many open items for completion of this protocol, is one of the topics addressed by ONE5G. The related DRX mechanisms were also described, as well the assumed UE power consumption reference model that ONE5G adopts. Proper usage of the many degrees of freedom that comes with the extended RRC protocol and DRX features is important for efficient leveraging of the latency, control plane signalling overhead, and UE power consumption.

In Chapter 3 we presented the preliminary work on different proposed solutions to radio resource management for improving the E2E performance 5G NR, including the study of RRC state handling and DRX, novel resource allocation techniques that account for the different requirements and properties of different services, and studies of signalling and control plane optimizations.

The work on optimized RRC state handling has so far established a preliminary framework for when which RRC state is to be used, whereas another contribution focuses on optimizing RRC parameters so as to improve performance metrics such as energy consumption, signalling overhead, etc. Also, another contribution studied different proposed strategies for switching Bandwidth Part (BWP) in DRX, considering performance impact in terms of e.g. signalling overhead and power consumption.

Thereafter, the chapter presents contributions that consider different perspectives of resource allocation optimization for multiple services. First, different enhancements of mitigating the damage caused to the eMBB traffic from using preemptive scheduling of URLLC was studied. Based on those, is concluded that preemptive scheduling is a promising technique, especially if combined with the proposed enhancements. Related contributions consider network slicing, both from the aspect of estimating the resources needed to satisfy the service requirements, as well as proposing algorithms for automating the negotiation of network slices and resources between operations. A number of contributions are studying specifically the challenges of C-RAN operation, specifically in terms of multi-cell scheduling and prediction based scheduling and flow optimization, in order to minimize interference impacts and leverage on different cross-cell cooperation schemes, enabled by C-RAN operation. A prerequisite for these schemes to work, is the collection of CQI for a large number of nodes. Another contribution therefore studies the efficient collection of CQI feedback. The last work item in this section outlines the intended work

on optimizing the scheduling of resources for mobile edge computing, which is a technique that allows mobile devices to offload computationally heavy tasks and save battery power.

One contribution concerning signalling and control plane optimizations, studies the effect of signalling overhead on the quality of the channel estimation, based on the compressive sensing approach. Another considered approach to optimizing signalling information, is to decouple data and control plane associations, so that optimal association criteria can be used for control and data planes, respectively, given the service type and network topology. Finally, the virtualization of mobile device capabilities in C-RAN and MEC is considered, and different architectures and functional split points will be studied.

In Chapter 4 we drafted different approaches to address an E2E-optimized multi-link management, including optimized usage of decoupled uplink and downlink cell associations, dynamic spectrum aggregation mechanisms, advanced mobility and load balancing techniques and device-to-device communications.

Specifically, the research on multi-link management has been split into a number of study items. First, a mechanism for the decision of the number of radio links and their usage (data split versus data duplication) in order to fulfill certain QoS requirements has been proposed by means of machine learning approaches. Then, a study on the decoupling of the rules for uplink and downlink cell association has been conducted, following a QoS-driven perspective in a highly heterogeneous network. Next, a study on the benefits in terms of reliability of data duplication via multiple connections has been performed, assuming uplink and downlink decoupling. And finally, the benefits and drawbacks of duplicating data at different levels (namely, the PDCP or the MAC layers) for reliability have been outlined.

The research on dynamic spectrum aggregation is conducted from different directions. On the one hand, it has been targeted through an evolution of MulteFire (the LTE standalone operation in the 5 GHz unlicensed band), in an attempt of reducing latency while improving reliability. On the other hand, a study on the QoS provided by the joint usage of licensed and unlicensed bands has been performed. Next, the benefits of dynamically adapting the bandwidth to the traffic demand in terms of UE power consumption have been assessed in the context of licensed bands, prior to its extension to unlicensed bands. Finally, the benefits in the increase of capacity have been derived in an ultra-dense urban environment, using both licensed and licensed bands.

Regarding mobility and load balancing enhancements, novel QoE-aware traffic steering mechanisms have been studied, including proactive techniques. Next, multi-access edge computing (MEC)-aware UE-cell connectivity has been studied, as a case of decoupled connectivity. And finally, load balancing has been addressed by means of time/space-variant network slicing.

As for D2D, different studies have been performed. First, schemes for device-to-device association and scheduling for interference reduction have been outlined, so as to maximize the network capacity in an eMBB environment. Then, relay-based schemes for coverage enhancement and reduction of power consumption in mMTC environments have been studied.

As stated for many of the presented techniques in this deliverable, the WP3 studies will continue, with final conclusions to appear in the next deliverable – D3.2. Furthermore, some contributions described in this deliverable will be potentially implemented and validated using the system level simulator developed in WP2. In fact, in collaboration with WP2, the techniques provided in Sections 3.2.5, 3.2.2, 4.1.3 and 4.3.1 are proposed for WP2 system level simulations.

Acknowledgment

Contribution from the following colleagues is also acknowledged: Aikaterini Demesticha, Aimilia Bantouna, Apostolos Voulkidis, Aspa Skalidi, Charilaos Kourogiorgas, Dimitrios Kelaidonis, Evangelia Tzifa, Ioannis Maistros, Konstantinos Tsoumanis, Kostas Trichias, Nelly Giannopoulou, Panagiotis Vlacheas, Paraskevas Bourgos, Yiouli Kritikou (WINGS).

References

[22.803] 3GPP TR 22.803 v1.0.0, “Technical Specification Group SA; Feasibility Study for proximity services”, Release 12, August 2012.

[23.303] 3GPP TS 23.303, “Proximity-based services (prose); stage 2”, technical Report, 2014.

[23.501] 3GPP 23.501, Technical Specification Group Services and System Aspects, “System Architecture for the 5G System – Stage-2”, Rel.15, September 2017.

[23.703] 3GPP TR23.703 V0.4.1, Technical Specification Group SA, “Study on architecture enhancements to support proximity services (ProSe)”, June 2013.

[36.842] 3GPP TR 36.842 v12.0.0, “Study on Small Cell enhancements for E-UTRA and E-UTRAN; Higher layer aspects”, Rel-12, December 2013.

[36.843] 3GPP TR 36.843, WG RAN1, “Feasibility Study on LTE Device to Device Proximity services – Radio Aspects”, March 2014.

[36.881] 3GPP TR 36.881 v14.0.0, “Study on latency reduction techniques for LTE”, Rel-14, June 2016.

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

[36.900] 3GPP TR 38.900 V14.2.0 (2016-12), “Study on channel model for frequency spectrum above 6 GHz”, Release14.

[37.324] 3GPP Technical Specification Group Radio Access Network, E-UTRA and NR, “Service Data Adaptation Protocol (SDAP) specification”, Rel-15, October 2017.

[37.340] 3GPP Technical Specification 37.340, “NR; Multi-connectivity; Overall description; Stage-2 (Rel-15)”, January 2018.

[38.211] 3GPP Technical Specification (TS) 38.211, ”NR; Physical Channels and Modulation”, Release 15, December 2017.

[38.300] 3GPP Technical Specification (TS) 38.300, ”NG Radio Access Network; Overall Description”, Stage 2, Release 15, September 2017.

[38.331] 3GPP Technical Specification (TS) 38.331, “Radio Resource Control (RRC); Protocol specification”, Release 15, November 2017.

[38.802] 3GPP Technical Report (TR) 38.802: "Study on New Radio (NR) Access Technology Physical Layer Aspects", March 2017.

[38.804] 3GPP Technical Report (TR) 38.804, “Study on New Radio Access Technology; Radio Interface Protocol Aspects (Release 14)”, March 2017.

[38.321] 3GPP Technical Specification 38.321, “NR; Medium Access Control (MAC) protocol specification (Rel-15)”, January 2018.

[38.323] 3GPP Technical Specification 38.323, “NR; Packet Data Convergence Protocol (PDCP) specification (Rel-15)”, January 2018.

[38.889] 3GPP TR38.889 “Study on NR-based Access to Unlicensed Spectrum”, Rel-15.

[38.901] 3GPP TR 38.901 V1.0.1 (2017-3) - 3rd Generation Partnership Project; Technical Specification Group Radio Access Network, “Study on channel model for frequencies from 0.5 to 100 GHz”, Release 14, 2017.

[38.913] 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, 2016.

[45.820] 3GPP, "Cellular system support for ultra low complexity and low throughput internet of things tech. rep. 45.820 v13.1.0," TSG GERAN, Tech. Rep., 2015.

[36.300] 3GPP Technical Specification 36.300, "Evolved Universal Terrestrial Radio Access (E-UTRA) and Evolved Universal Terrestrial Radio Access Network (E-UTRAN); Overall description; Stage 2," (Rel-15), September 2017.

[R1-1711465] Huawei, HiSilicon, "NR Numerology on unlicensed bands," R1-1711465, Qingdao, China, June 27-30, 2017.

[R1-1711467] Huawei, HiSilicon, "Coexistence and channel access for NR-based unlicensed band operation," R1-1711467, Qingdao, China, June 27-30, 2017.

[R1-1711469] Huawei, HiSilicon, "NR standalone operation on unlicensed bands,"R1-1711469, Qingdao, China, June 27-30, 2017.

[RP-162235] "Core part RP-152272, perf. part RP-162235".

[RP-172021] 3GPP RP-172021, "Study on NR-based Access to Unlicensed Spectrum", September 2017; available at www.3gpp.org

[RP-170295] 3GPP RP-170295, "Study on Further Enhancements to LTE Device to Device, UE to Network Relays for IoT and Wearables, » March 2017; available at www.3gpp.org

[R2-1709220] 3GPP Technical Document R2-1709220, "Harmonization of the RRC procedures", Nokia, August 2017.

[R2-168299] 3GPP, "How to realize zero ms UP interruption in NR", Ericsson, November 2016.

[AWM+14] A. Asadi, Q. Wang and V. Mancuso, "A Survey on Device-to-Device Communication in Cellular Networks," in IEEE Communications Surveys and Tutorials, vol. 16, no. 4, pp. 1801-1819, Fourthquarter 2014.

[BAL+17] N. Baldo, M. Miozzo, M. Requena-Esteso and J. Nin-Guerrero, "LENA (LTE-EPC network simulator) module for ns-3", [online] available at: https://www.nsnam.org/docs/models/html/lte-design.htmlhttps://www.nsnam.org/docs/models/html/lte-design.html

[BB15] H. Boostanimehr and V. K. Bhargava, "Unified and Distributed QoS-Driven Cell Association Algorithms in Heterogeneous Networks," in IEEE Transactions on Wireless Communications, vol. 14, no. 3, pp. 1650-1662, March 2015.

[BEC+12] H. Becker, D. Iter, M. Naaman, and L. Gravano, "Identifying content for planned events across social media sites," in Proceedings of the Fifth ACM International Conference on Web Search and Data Mining, ser. WSDM '12. New York, NY, USA: ACM, 2012, pp. 533–542. [Online]. Available: http://doi.acm.org/10.1145/2124295.2124360.

[CWP16] Cisco White Paper, "Cisco visual networking index: Forecast and methodology, 2015-2020", Technical Report, June 2016. [Online] available at: http://www.cisco.com/c/en/us/solutions/collateral/service-provider/ip-ngn-ip-next-generation-network/white paper c11-481360.pdf.

[CWP17] Cisco White Paper, "Cisco visual networking index: Global mobile data traffic forecast update, 2016-2021", Tech. Rep., February 2017.

[D21] ONE5G Deliverable D2.1, “Scenarios, KPIs, use cases and baseline system evaluation”, 2017.

[DAD+17] M. Deghel, M. Assaad, M. Debbah, A. Ephremides, “Traffic-Aware Scheduling and Feedback Allocation in Multichannel Wireless Networks”, Provisionally accepted in IEEE Transactions on Wireless Communications, 2017.

[DAD+18] M. Deghel, M. Assaad, M. Debbah, A. Ephremides, “Queueing Stability and CSI Probing of a TDD Wireless Network with Interference Alignment,” in IEEE Transactions on Information Theory, (64):1, pp. 547-576, 2018.

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

[ECPRI17] eCPRI Transport Network D0.1, 30 August 2017. Available for download at: http:/www.cpri.info/downloads/Requirements_for_the_eCPRI_Transport_Network_d_0_1_2017_08_30.pdf

[EFS17] M. Emara, M. Filippou and D. Sabella, “MEC-aware Cell Association for 5G Heterogeneous Networks,” in 2018 IEEE Wireless Communications and Networking Conference Workshops, April 2018.

[EFS18] M. Emara, M. Filippou and D. Sabella,“MEC-assisted End-to-End Latency Evaluations for C-V2X Communications,” in 2018 European Conference on Networks and Communications (EuCNC) (to appear).

[EP18] A. Esswie, K.I. Pedersen, ”Multi-User Preemptive Scheduling For Critical Low Latency Communications in 5G Networks”, accepted for publication in IEEE ISCC, June 2018.

[Fan5G D4.2] FANTASTIC5G, Deliverable 4.2, “Final Results for the flexible 5G Air Interface multi-node/multi-antenna solution”, April 2017.

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

[FDS+18] S. Fortes, D. Palacios, I. Serrano, R.Barco, “Applying Social Event Data for the Management of Cellular Networks”, provisionally accepted in IEEE Communications Magazine, 2017

[FJH+16] F. Boccardi et al., "Why to decouple the uplink and downlink in cellular networks and how to do it," in IEEE Communications Magazine, vol. 54, no. 3, pp. 110-117, March 2016

[FZY16] C. Fan, Y. J. Zhang, and X. Yuan, “Dynamic nested clustering for parallel PHY-layer processing in cloud-RANs,” IEEE Trans. Wireless Commun., vol. 15, no. 3, pp. 1881–1894, Mar. 2016.

[GJK+16] Guowang Miao, Jens Zander, Ki Won Sung, and Sliman Ben Sliman, *Fundamentals of Mobile Data Networks*, Cambridge University Press, 3 Mar 2016.

[GMP16] L. C. Giménez, P. H. Michaelsen and K. I. Pedersen, "UE autonomous cell management in a high-speed scenario with dual connectivity," *2016 IEEE 27th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC)*, Valencia, 2016, pp. 1-6.

[GMP+17] L. C. Gimenez, P. H. Michaelsen, K. I. Pedersen, T. E. Kolding and H. C. Nguyen, "Towards Zero Data Interruption Time with Enhanced Synchronous Handover," *2017 IEEE 85th Vehicular Technology Conference (VTC Spring)*, Sydney, NSW, 2017, pp. 1-6.

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

[Hae16] M. Haenggi, "The Meta Distribution of the SIR in Poisson Bipolar and Cellular Networks," in IEEE Transactions on Wireless Communications, vol. 15, no. 4, pp. 2577-2589, April 2016.

[HFM+14] H. Elshaer, F. Boccardi, M. Dohler, and R. Irmer, “Downlink and uplink decoupling: A disruptive architectural design for 5G networks,” in 2014 IEEE Global Communications Conference, Dec 2014, pp. 1798–1803

[HSV16] B. Héder, P. Szilágyi, C. Vulkán, ”Dynamic and Adaptive QoE Management for OTT Application Sessions in LTE”, in IEEE Proc. International Symposium on Personal Indoor and Mobile Radio Communications (PIMRC), September 2016.

[ITU2015] ITU Recommendation M.2083, “IMT Vision – Framework and overall objectives of the future development of IMT for 2020 and beyond”, Sept. 2015.

[JPL+01] N. R. Jennings, P. Faratin, A. R. Lomuscio, S. Parsons, C. Sierra and M. Wooldridge (2001) "Automated negotiation: prospects methods and challenges" Int. J. of Group Decision and Negotiation 10 (2) 199-215.

[KW18] M. Kasparick, G. Wunder, “Stable Wireless Network Control Under Service Constraints,” IEEE Transactions on Control of Network Systems, 2018, to appear, [online]: https://arxiv.org/abs/1701.04201

[KPL15] K. Smiljkovikj, P. Popovski and L. Gavrilovska, "Analysis of the Decoupled Access for Downlink and Uplink in Wireless Heterogeneous Networks," in IEEE Wireless Communications Letters, vol. 4, no. 2, pp. 173-176, April 2015.

[LAH04] Haifei Li, D. Ahn and P. C. K. Hung, "Algorithms for automated negotiations and their applications in information privacy," *Proceedings. IEEE International Conference on e-Commerce Technology, 2004. CEC 2004.*, 2004, pp. 255-262

[LAK17] H. Q. Le, H. Al-Shatri, and A. Klein, “Efficient resource allocation in mobile-edge computation offloading: Completion time minimization,” in 2017 IEEE International Symposium on Information Theory (ISIT), June 2017, pp. 2513–2517.

[LBY+15] Y. Lin, W. Bao, W. Yu and B. Liang, "Optimizing User Association and Spectrum Allocation in HetNets: A Utility Perspective," in IEEE Journal on Selected Areas in Communications, vol. 33, no. 6, pp. 1025-1039, June 2015.

[LBL+16] Q. Liao, P. Baracca, D. Lopez-Perez, L.G. Giordano, “Resource Scheduling for Mixed Traffic Types with Scalable TTI in Dynamic TDD Systems”, in IEEE Proc. Globecom, December 2016.

[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.

[LVM17] H. Lee, S. Vahid, and K. Moessner, “Traffic-aware carrier allocation with aggregation for load balancing,” in 2017 European Conference on Networks and Communications (EuCNC), 2017, pp. 1–6.

[LEM16] M. A. Lema, E. Pardo, O. Galinina, S. Andreev, and M. Dohler, “Flexible Dual-Connectivity Spectrum Aggregation for Decoupled Uplink and Downlink Access in 5G Heterogeneous Systems,” IEEE Journal on Selected Areas in Communications, vol. 34, no. 11, pp. 2851–2865, Nov. 2016.

[LLR+18] D. Laselva, D. Lopez-Perez, M. Rinne and T. Henttonen, “3GPP LTE-WLAN Aggregation Technologies: Functionalities and Performance Comparison,” in *IEEE Communications Magazine*, vol. 56, no. 3, pp. 195-203, March 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.

[MAB+16] A. Maeder, A. Amaanat, A. Bedekar, A. Cattoni, D. Chandramouli, S. Chandrahekar, L. Du, M. Hesse, C. Sartori, S. Turtinen, "A Scalable and Flexible Radio Access Network Architecture for Fifth Generation Mobile Networks", in IEEE Communications Magazine, vol. 54, no. 11, pp. 16-23, November 2016.

[MAH16] G. Miao, A. Azari, and T. Hwang, "E2-MAC: Energy efficient medium access for massive M2M communications," IEEE Trans. Commun.,vol. 64, no. 11, pp. 4720-4735, Nov. 2016.

[Mey07] S. Meyn, Control Techniques for Complex Networks, Cambridge University Press New York, NY, USA, 2007.

[MNK+2016] G. C. Madueño, J. J. Nielsen, D. M. Kim, N. K. Pratas, Č. Stefanović and P. Popovski, "Assessment of LTE Wireless Access for Monitoring of Energy Distribution in the Smart Grid," in IEEE Journal on Selected Areas in Communications, vol. 34, no. 3, pp. 675-688, March 2016.

[NGMN15] NGMN Alliance 5G White Paper, 17 February 2015, available for download at: https://www.ngmn.org/fileadmin/ngmn/content/downloads/Technical/2015/NGMN_5G_White_Paper_V1_0.pdf

[MULT] https://www.multefire.org/

[MULTR1] "MulteFire Release 1.0 Technical Paper: A New Way to Wireless", available at www.multefire.org

[MZL17] Y. Mao, J. Zhang, and K. B. Letaief, "Joint task offloading scheduling and transmit power allocation for mobile-edge computing systems," in 2017 IEEE Wireless Communications and Networking Conference (WCNC), March 2017, pp. 1–6.

[Nee03] M. Neely, "Dynamic power allocation and routing for satellite and wireless networks with time varying channels » Ph.D. Dissertation, MIT, 2003.

[OLI+16] P. Oliver-Balsalobre, M. Toril, S. Luna-Ramírez, and J. M. R. Avilés: Self-tuning of scheduling parameters for balancing the quality of experience among services in LTE, EURASIP Journal on Wireless Communications and Networking, (2016). vol. 7, pp. 1–12.

[OY13] M. Ouyang and L. Ying, "Approaching throughput optimality with limited feedback in multichannel wireless Downlink networks," IEEE/ACM Transactions on Networking, vol. 21, no. 6, pp. 1827–1838, Dec 2013.

[PBF+17] K.I. Pedersen, G. Berardinelli, F. Frederiksen, and P. Mogensen, "A Flexible 5G Wide Area Solution for TDD with Asymmetric Link Operation", IEEE Wireless Communications, vol. 24, no. 2, pp. 122–128, April 2017.

[PBM11] G. Piro, N. Baldo, and M. Miozzo, "An LTE module for ns-3 network simulator," Proc. of Int. ICST Conf. on Simulation Tools and Techniques, Mar. 2011.

[PKZ16] J. Park, S.-L. Kim, and J. Zander, "Tractable resource management with uplink decoupled millimeter-wave overlay in ultra-dense cellular networks," IEEE Trans. Wireless Commun., vol. 15, no. 6, pp. 4362–4379, Jun. 2016.

[PNS+16] K.I. Pedersen, M. Niparko, J. Steiner, J. Oszmianski, L. Mudolo, S.R. Khosravirad, "System Level Analysis of Dynamic User-Centric Scheduling for a Flexible 5G Design", in IEEE Proc. Globecom, December 2016.

[PPS+16] G. Pocovi, K.I. Pedersen, B. Soret, M. Lauridsen, P.E. Mogensen, ”On the Impact of Multi-User Traffic Dynamics on Low Latency Communications”, in Proc. International Symposium on Wireless Communication Systems (ISWCS), September 2016.

[PPS+17] K.I. Pedersen, G. Pocovi, J. Steiner, S. Khosravirad, “Punctured Scheduling for Critical Low Latency Data on a Shared Channel with Mobile Broadband”, in IEEE Proc. VTC-fall, September 2017.

[PPS18] K.I. Pedersen, G. Pocovi, J. Steiner, ” Pre-emptive Scheduling of Latency Critical Traffic and its Impact on Mobile Broadband Performance”, accepted, VTC-2018-spring (recent results). June 2018

[PPS+18] K. Pedersen, G. Pocovi, J. Steiner, A. Maeder, ”Agile 5G Scheduler for Improved E2E Performance and Flexibility for Different Network Implementations”, in IEEE Communications Magazine, March 2018.

[PSP+17] G. Pocovi, B. Soret, K.I. Pedersen, P.E. Mogensen, “MAC Layer Enhancements for Ultra-Reliable Low-Latency Communications in Cellular Networks”, in IEEE Proc ICC (workshop), June 2017.

[PRB+11] P. Munoz, R. Barco, I. de la Bandera, M. Toril and S. Luna-Ramirez, "Optimization of a Fuzzy Logic Controller for Handover-Based Load Balancing," 2011 IEEE 73rd Vehicular Technology Conference (VTC Spring), Budapest, 2011, pp. 1-5.

[PRL+13] P. Munoz, R. Barco, D. Laselva and P. Mogensen, "Mobility-based strategies for traffic steering in heterogeneous networks," in IEEE Communications Magazine, vol. 51, no. 5, pp. 54-62, May 2013.

[RAN1-91bis] 3GPP RAN1-91bis, Chairman notes, October 2017.

[ROS16] C. Rosa et al., “Dual connectivity for LTE small cell evolution: functionality and performance aspects,” IEEE Communications Magazine, vol. 54, no. 6, pp. 137–143, Jun. 2016.

[RPW+16] C. Rosa *et al*., "Dual connectivity for LTE small cell evolution: functionality and performance aspects," in *IEEE Communications Magazine*, vol. 54, no. 6, pp. 137-143, June 2016.

[SKD+05] V. Stavroulaki, A. Katidiotis, G. Dimitrakopoulos, P. Demestichas and S. Buljore, "Negotiation and selection of equipment reconfigurations in beyond 3G systems," *2005 IEEE 16th International Symposium on Personal, Indoor and Mobile Radio Communications*, Berlin, 2005, pp. 1969-1973 Vol. 3

[SW18] S. Stefanatos and G. Wunder, “Performance Limits of Compressive Sensing Channel Estimation in Dense Cloud RAN,” ICC 2018, accepted, [online]: https://arxiv.org/pdf/1710.10796.

[SW18b] R. Schoeffauer and G. Wunder, “Model Predictive Network Control and Throughput Sub-Optimality of MaxWeight,” submitted to EuCNC 2018, [online]: https://arxiv.org/abs/1804.00481.

[TA92] L. Tassiulas and A. Ephremides, “Stability Properties of Constrained Queueing Systems and Scheduling Policies for Maximum Throughput in Multihop Radio Networks,” IEEE Transactions on Automatic Control, vol. 31, no. 12, Dec. 1992.

[TES16] F. B. Tesema, A. Awada, I. Viering, M. Simsek, and G. P. Fettweis, “Evaluation of adaptive active set management for multi-connectivity in intra-frequency 5G networks,” in 2016 IEEE Wireless Communications and Networking Conference, 2016, pp. 1–6.

[VAV+18] C. Vargas Anamuro, N. Varsier, J. Schwoerer and X. Lagrange, “Simple modeling of energy consumption for D2D relay mechanism,” WCNC 2018 CmMmW5G workshop, accepted.

[WAN10] Y. Wang, K. I. Pedersen, T. B. Sorensen, and P. E. Mogensen, “Carrier load balancing and packet scheduling for multi-carrier systems,” IEEE Transactions on Wireless Communications, vol. 9, no. 5, pp. 1780–1789, May 2010.

[WBS+15] G. Wunder, H. Boche, T. Strohmer, and P. Jung, “Sparse signal processing concepts for efficient 5G system design,” IEEE Access, vol. 3, pp.195–208, 2015.

[WRC19] http://www.itu.int/en/ITU-R/conferences/wrc/2019/Pages/default.aspx

6 Appendix

6.1 Annex A: RRC Design for Multiple Network Slices

Assuming Poisson process for event based traffic with mean arrival rate $\lambda=1/T_{ini}$ as shown in Figure 3-1(a), we consider the general case where N packets are assumed to arrive between time $(l-1)T_p$ and lT_p The n-th packet arrives at time

$$s_n = (l-1)T_p + \sum_{i=1}^{n} x_i, \tag{1}$$

where x_i is the inter-arrival time between packet (i-1) and i except x_1. Since all the packets arrived during time [(l-1)T_p, lT_p) are buffered at the eNB and sent until time lT_p, the delay of the n-th packet is given as

$$d_n = lT_p - (l-1)T_p - \sum_{i=1}^{n} x_i = T_p - \sum_{i=1}^{n} x_i. \tag{2}$$

For the first packet, the inter-arrival time between itself and the previous packet is z_1. Clearly, z_1 and x_2 to x_N follow the exponential distribution and are independent from each other. If l=1, z_1=x_1; otherwise $z_1 \geq x_1$. However, x_1 also follows the exponential distribution and is independent from x_2 to x_N. This means that d_n has no relevance to l so that we can consider a simplified case as shown in Figure 3-1(b) and we have

$$s_n = \sum_{i=1}^{n} x_i, \tag{3}$$

where s_n follows the Erlang distribution.

If N=1, i.e., there is only one packet arrived during time [0, T_p). The joint density for X_1 and S_2 is

$$f_{X_1 S_2}(x_1, s_2) = f_{X_1}(x_1) f_{X_2}(s_2 - x_1). \tag{4}$$

The marginal density of S_2 can be obtained by integrating X_1 out from the joint density, which takes the form:

$$\begin{aligned} f_{X_1 S_2}(x_1, s_2) &= \lambda^2 \exp(-\lambda x_1) \exp(-\lambda (s_2 - x_1)) \\ &= \lambda^2 \exp(-\lambda s_2), \quad \text{for } 0 \leq x_1 \leq s_2. \end{aligned} \tag{5}$$

6.2 Annex B: Simulation parameters related to the Reliability Oriented MC Technique

System Level Simulation parameters for the preliminary results presented in Section 4.1.6. The simulation assumptions are shown in Table 6-1. The focus is on downlink evaluation only. There is a 15 kHz spacing between subcarriers and Orthogonal Frequency Division Multiple Access (OFDMA) is used to dynamically multiplex users on a shared channel. A short TTI of 2 OFDM symbols is used (0.143 ms). In the frequency domain, a Physical Resource Block (PRB) resolution of 12 subcarriers can be used to multiplex UEs. Asynchronous Hybrid Automatic Repeat Request (HARQ) is used. A BLER target of 1% is set.

Table 6-1: System Level Simulation Parameters.

Parameters	Macro layer-NR	Pico layer -- NR
Layout	7 sites, 21 cells, wrap around	4 picos per cell
Inter-BS distance	500m	cluster
Carrier frequency	2 GHz	3.5 GHz
Simulation bandwidth	10 MHz	
BS power	46 dBm	30 dBm
Pathloss Model	3D-UMa	3D-UMi
Antenna Height	32m	10m
UE Antenna Height	1.5m	
Antenna gain	18dBi	5dBi
UE Antenna gain	0dBi	
Antenna configuration	2x2 cross-polar	
UE dropping	2/3 UEs randomly and uniformly dropped within the clusters, 1/3 UEs randomly and uniformly dropped throughout the macro	
Radius for small cells dropping in a cluster	50m	
Radius for UE dropping in a cluster	70m	
Minimum distance	Small Cell – Small Cell: 20m Small Cell – UE: 5m	

(2D Distance)	Macro – Small Cell Cluster Centre: 105m Macro – UE: 35m Cluster Centre – Cluster Centre: 100m	
BS antenna pattern	TR36_814	Isotropic
BS antenna height	32 m	10 m
Subcarriers/RB	12	
Subcarrier spacinSswg(kHz)	15	
Cell Selection Criteria	RSRP with CRE	
Receiver Type	LMMSE_IRC	
Traffic Model	FTP3: based on FTP model 2 with the exception that packets for the same UE arrive according to a Poisson process and the transmission time of a packet is counted from the time instance it arrives in the queue. Payload = 32 bytes	
LLC Transport Type	UDP	
Maximum PDU Size	1500B	
Link Adaptation	Outer Loop Link Adaptation (OLLA) algorithm 1% initial BLER target for LLC	

6.3 Annex C: Efficient CQI Scheduling

In this section, we provide a description of the work summarized in Section 3.2.7. For more details one can refer to [DAD+17].

System Model and Proposed Solution

We consider an FDD cellular wireless network, with one single-antenna BS, N single-antenna mobile users and L RBs. The packets to be transmitted to the users are stored in N separate queues at the BS. Let $q_i(t)$ denote the length of queue i at the beginning of time-slot t. The channel state of a user on a RB represents the bit rate such that the paclets are succesfully received. We use $C_{ij}(t)$ to represent the state of link (ij) (i.e. user I and RB j) at time-slot t. We assume that each link state can take K possible values $\{R_1,...,R_K\}$, where rate R_k is used if the SNR of the link (ij) is between two threshold values, i.e. . $\tau_k \le \gamma_{ij}(t) < \tau_{k+1}$. We suppose that the values R_k are sorted in a descending order. Furthermore, we consider that $C_{ij}(t)$ varies from one time slot to another due to channel fading, which is modeled here as a *channel convergent* model [Nee03]. Note that this general model is widely used to model a Markovian channel since in general the time varying channel is not necessarily i.i.d. over slots. We consider a realistic context where the CQI knoweldge is delayed and the feedback is limited. This can be modeled as follows. We use $\hat{C}_{ij}(t)$ to denote the rate of link (ij) at time-slot t-d, i.e. $\hat{C}_{ij}(t)=C_{ij}(t-d)$.

Let us define $S_{ij}(t)$ to be the scheduling decision associated with link (ij) at time-slot t. In addition, we define $Y_{ij}(t)$ as the feedback decision associated with link (ij) at time-slot t. Recall that, due to the feedback delay, the feedback decision is made d slots before the transmission of the data. Let $\hat{Y}_{ij}(t)=Y_{ij}(t-d)$ denote the feedback decision associated with link (ij) at time-slot t-d. So, we can write

$$\hat{Y}_{ij}(t)=\begin{cases} 1, & \text{if } \hat{C}_{ij}(t) \text{ gets reported to the scheduler} \\ 0, & \text{otherwise,} \end{cases}$$

where we recall that $\hat{C}_{ij}(t)=C_{ij}(t-d)$. On the other side, for $S_{ij}(t)$, which represents the scheduling decision at time-slot t, we have

$$S_{ij}(t)=\begin{cases} 1, & \text{if user } i \text{ gets scheduled on channel } j \\ 0, & \text{otherwise.} \end{cases}$$

The queueing dynamics are then given as follows

$$q_i(t+1)=\max\{q_i(t)+A_i(t)-\sum_{j=1}^{L}\hat{C}_{ij}(t)\hat{Y}_{ij}(t)S_{ij}(t)1_{(C_{ij}(t)\ge\hat{C}_{ij}(t))},0\},\text{for } 1\le i\le N,$$

where the expression

$$\sum_{j=1}^{L}\hat{C}_{ij}(t)\hat{Y}_{ij}(t)S_{ij}(t)1_{(C_{ij}(t)\ge\hat{C}_{ij}(t))}$$

is the service rate allocated for user i at time-slot t. The use of indicator function $1_{(C_{ij}(t)\ge\hat{C}_{ij}(t))}$ is due to the outage that may occur when the actual channel conditions cannot support the reported rate, $\hat{C}_{ij}(t)$.

Our objective here is to develop a feedback and scheduling strategy that stabilizes the queues of the users whenever it is possible. In addition, we evaluate the performance gap between our policy and the ideal system (i.e. the system in which the CQIs for all users and RBs are available at each

time at the BS at no cost). Before describing our proposed policy, we will first provide the definitions of strong stability and stability region of a queueing system.

Definition 1 (Strong Stability) *The condition for strong stability can be expressed as*

$$\limsup_{T\to\infty}\frac{1}{T}\sum_{t=0}^{T-1}\mathsf{E}\left\{q_i(t)\right\}<\infty,\forall i\in\{1,...,N\}.$$

Definition 2 (Stability Region) *The stability region achieved by a scheduling policy is defined as the set of vectors of mean arrival rates for which the queues stay strongly stable under this policy.*

Proposed Algorithm and its Stability Performance

1. Queue lengths broadcast every T_b slots:

Every T_b time-slots, the BS broadcasts the queue lengths of all users. Each user has then an outdated knowledge of the state of the queues. Let $\tilde{q}_i(t)$ be the (outdated) queue length the users know at time t, i.e. $\tilde{q}_i(t)=q_i(nT_b)$.

2. *Feedback and scheduling decisions at time-slot t-d:*

Let us denote $\tilde{q}_i(t-d)$ as $\hat{\tilde{q}}_i(t)$. For each RB, only one user sends its CSI to the BS. This is done by letting the users contend among each other as follows: User i, waits until time $T_c\left(\hat{\tilde{q}}_i(t)\hat{C}_{ij}(t)\mathsf{P}\left\{C_{ij}(t)\geq\hat{C}_{ij}(t)\,|\,\hat{h}_{ij}(t)\right\}\right)^{-1}$, and then broadcasts a signal (of negligible duration) to end the contention procedure of RB j. Note that T_c is the contention period for each RB. The corresponding user reports its CQI. Then, the contention of another RB gets started.

3. *Transmission at time-slot t:*

At the end of the contention period of all RBs, the BS has the CQI of each RB. The user selected to report its CQI of RB j will be scheduled for data transmission on this RB.

We now provide the stability region that our policy (denoted by Λ_{mdl}) can achieve compared with the stability region of the ideal system.

Theorem *Our policy achieves at least a fraction β of the stability region achieved by the ideal system, i.e. the region Λ_{mdl} can be bounded as*

$$\beta\Lambda_{\text{pf}}\subseteq\Lambda_{\text{mdl}}\subseteq\Lambda_{\text{pf}},$$

where $\beta=(1-\frac{1}{T_b})\frac{p_c^{\min}}{\eta}$, $\eta=\max_{t,(ij)}\left\{\frac{C_{ij}(t)}{\hat{C}_{ij}(t)}\right\}$, $p_{cij}^{\min}=\min_{(ij)}\min_{t,\hat{h}_{ij}(t)}\left\{\mathsf{P}\{C_{ij}(t)\geq\hat{C}_{ij}(t)\,|\,\hat{h}_{ij}(t)\}\right\}$.

The proof of the aforementioned theorem can be found in our paper in [DAD+17]. As future work, other feedback strategies and more performance analysis will be investigated. Simulations, taking into account the Use cases defined in WP2, will be provided.

6.4 Annex D: Performance Analysis of D2D Networks with Interference Alignment

In this section, we provide more details on the system model and the main results of the work presented in section 4.4.1..

System Model

We consider the MIMO interference channel with N transmitter-receiver pairs shown in Figure 4-29. We assume that all transmitters are equipped with N_t antennas and all receivers with N_r antennas. Only a subset of pairs, of cardinality $L(t)$, is active at time-slot t. Transmitter k has d_k independent data streams to transmit to its intended user k. For the cases where $L(t)$ is higher than 2, while each transmitter communicates with its intended receiver, it also creates interference to other unintended receivers.

We assume a TDD transmission strategy, which enables the transmitters to estimate their channels toward different users by exploiting the reciprocity of the wireless channel. As alluded earlier, global CSI is required at each transmitting node in order to design the IA. Each transmitter k has then to quantize its local estimated channels H_{ik} (where i is the receiver index) and to exchange them among each other over the backhaul. The quantization is done using a codebook with a number of quantization bits equal to B. This quantization has certainly an impact on the accuracy of the assigned rate to each transmitter k. In this work, we adopt the following rate model. At each time, the instantaneous rate R of stream m of transmitter k is transmitted successfully if the SINR $\gamma_k^{(m)}$ at the corresponding receiver is higher than or equal to a given threshold τ. Furthermore, we denote by $\widetilde{R}_k(t)$ the assigned rate for user k at time t, thus $\widetilde{R}_k(t)$ is the sum of the assigned rates for all the streams of transmitter k at time t. In other words, we have

$$\tilde{R}_k(t) = \sum_{m=1}^{d_k} R \, 1_{\left(\gamma_k^{(m)}(t) \geq \tau\right)},$$

where $1_{(\cdot)}$ is the indicator function. As mentioned earlier, in this work the channel acquisition cost is taken into account. Since we consider a system under TDD mode, the local CSIs are acquired by each transmitter by decoding the pilots sent by the users. This assumes that the users use orthogonal sequences with lengths obviously proportional to the number of active users. Transmitting the pilots of $L(t)$ active users requires a fraction $L(t)\theta$ of the time-slot. The *actual* rate $D_k(t)$ at time-slot t is

$$D_k(t) = (1 - L(t))\tilde{R}_k(t)$$

For each user, we assume that the incoming data is stored in a respective queue (i.e. buffer) until transmission, and we denote by $\mathbf{q}(t)=(q_1(t),\ldots,q_N(t))$ the queue length vector. We designate by $A_k(t)$ the number of bits arriving in the buffer of transmitter k in time-slot t, which is an i.i.d. in time process and independent across users. The queue length is then given by,

$$q_k(t+1) = \max\{q_k(t) - D_k(t), 0\} + A_k(t)$$

The mean arrival rate (in units of bits/slot) for user k is denoted by a_k. At each time-slot, a set of pairs is scheduled (of cardinality L(t)), based on the queue lengths and the average rates in the system. The selected active users send their pilots in the uplink so that the (active) transmitters can estimate the CSI. The decision of selecting active pairs is referred as the *scheduling policy*. At the t-th slot, this policy can be represented by an indicator vector $s(t) \in \{0,1\}^N$, where the k-th component of s(t), denoted $s_k(t)$, is equal to 1 if the k-th queue is scheduled or otherwise equal to 0.

In this work, the focus will be mainly on the stability of the system. Thus, in the following we provide the definitions of "Strong Stability" and "Stability region" of a system.

Definition 1 (Strong Stability) *The condition for strong stability of the system can be expressed as*

$$\limsup_{T\to\infty}\frac{1}{T}\sum_{t=0}^{T-1}\sum_{i}^{N}\mathsf{E}\left\{q_i(t)\right\}<\infty,\forall i\in\{1,...,N\}.$$

Definition 2 (Stability Region) *The stability region of the aforementioned model is defined as the set of vectors of mean arrival rates for which the system is strongly stable. Furthermore, a scheduling policy that achieves this region is called throughput optimal.*

In this work, we will characterize the stabiltiy region of the aforementioned network model when IA and SVD are employed taking into account the limited capacity of the backhaul.

Performance Analysis

We first define subset S_L as the subset of scheduling decision vectors for which the number of active pairs is equal to L. The subset of average rate vectors is defined as $I_L=\{r(L)s: s\in S_L\}$. For L=1, when the SVD technique is used, the subset of average rate vectors is defined as $I_1=\{r_{svd}s: s\in S_1\}$. We define set $I=\{0,I_1,I_2,\dots,I_{L_I}\}$, i.e. it contains the origin point 0 and the set of average rate vectors when the number of active pairs L is between 1 and L_I. We also define set $\bar{I}=\{I_{L_I+1,\dots},I_N\}$, i.e. it contains the set of average rate vectors for which L is between L_I+1 and N. Using these definitions, we can state the theorem to characterize the stability region of the system.

Theorem *The stability region of the system employing IA with limited backhaul can be characterized as* $\Lambda_I=CH\{0,I_1,I_2,\dots,I_{L_I}\}$ where CH represents the convex hull.

The above theorem means that the stability region can be characterized by a scheduling policy that activates at each time a subset L_I of the pairs. By increasing the number of active pairs L beyond L_I, the set of achievable average rates will not increase. One can refer to [DAD+18] for more details on the proof of the theorem and the exact characterization of L_I.

We now provide the conditions under which IA yields a queueing stability gain with respect to SVD. These conditions are given as follows.

Proposition *IA provides a queueing stability gain iff there exists a number L such that* $Lr(L)>r_{svd}$ *, with* $2\le L\le L_I$ *, where*

$$r(L)=(1-L\theta)dRe^{-\frac{\sigma^2\tau}{\alpha}}F^{L-1},$$

$$r_{\mathrm{svd},k}=(1-\theta)d_kR\sum_{n=0}^{m_2-1}\Omega_n\sum_{j=0}^{2n}\kappa_j\Gamma(j+m_1-m_2+1,\frac{d_k\sigma^2\tau}{\zeta_{kk}P}),$$

in which

$$F=\left(d\tau 2^{-\frac{B}{Q}}+1\right)^{-Q}{}_2F_1\left(c_2,Q;c_1+c_2;(2^{\frac{B}{Q}}(d\tau)^{-1}+1)^{-1}\right).$$

$$\Omega_n=n!\left(m_2(n+m_1-m_2)!\right)^{-1},\ \kappa_j=\sum_{i=0}^{j}\omega_i\omega_{j-i},$$

$$\omega_l=(-1)^l(n+m_1-m_2)!\left((n-l)!(m_1-m_2+l)!l!\right)^{-1}.$$

where $Q=N_tN_r-1$, ${}_2F_1$ is the Hypergeometric function, $c_1=(Q+1)Q^{-1}d-Q^{-1}$ and $c_2=(Q-1)c_1$. $\Gamma(.,.)$ is the upper incomplete Gamma function. P is the total power at each transmitting node, σ^2 is the noise variance, and ξ_{ki} represents the path loss of channel H_{ki}. More details on the proof and interpretation of the above results can be found in [DAD+18].

6.5 Annex E: Minimizing the global energy consumption in relaying mode

For the first analysis only Tx active and Rx active states are considered for both MTD and UE. We study two transmission modes: when an MTD uses a direct transmission to the BS (direct mode) and when it uses a relaying mechanism (relaying mode). Figure 6-1 shows the relaying process in terms of power consumption during Tx active and Rx active states and duration of these states.

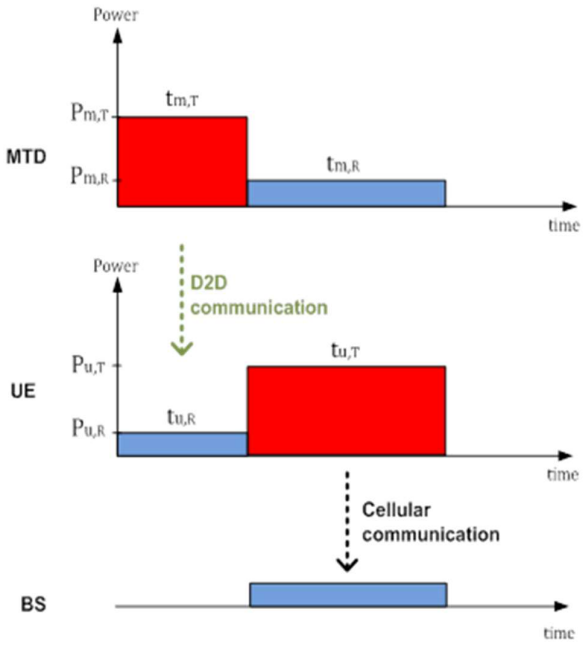


Figure 6-1: Relaying process

We are looking for minimizing the global energy consumption (energy consumption by the MTD + energy consumption by the UE). The minimum global energy consumption is obtained when the BS, the MTD and the UE are aligned (Figure 6-2)

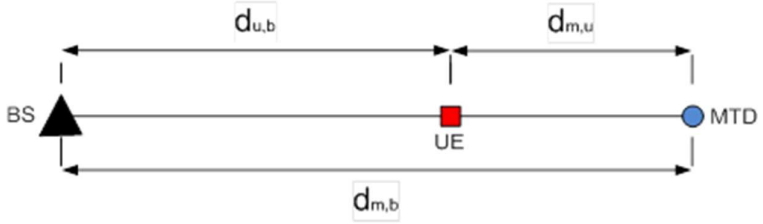


Figure 6-2: Minimum global energy consumption configuration

Figure 6-3 shows the global energy consumption in direct versus relaying mode as a function of the normalized distance between the BS and the UE and the distance between the BS and the MTD considering that the BS, the UE, and the MTD are aligned.

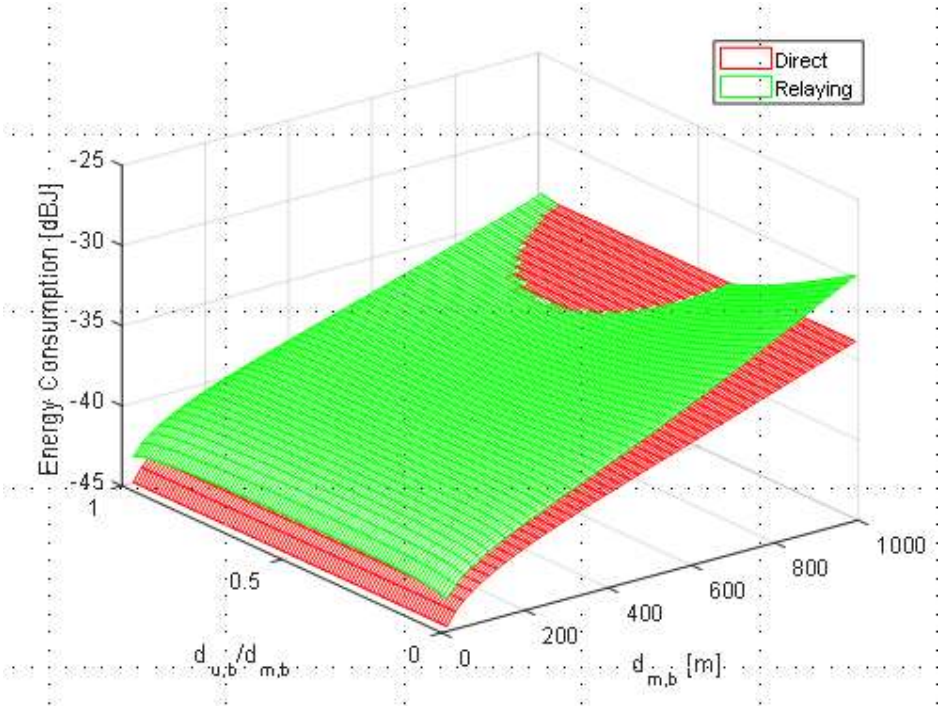


Figure 6-3: Global energy consumption in direct versus relaying mode as a function of the normalized distance between the BS and the UE and the distance between the BS and the MTD considering that the BS, the UE, and the MTD are aligned

For a simple energy consumption model, depending on the distance between the MTD, the relay, and the BS, we show that when the MTD is close to the base station, D2D relay mechanism can reduce the energy consumed by the MTD at the expense of an increase in the global energy consumption (i.e. the energy consumed by the MTD plus the energy consumed by the relay). On the other hand, when the MTD is far from the base station, D2D relay mechanism allows reducing both the energy consumption by the MTD and the global energy consumption. More details on the interpretation of the above results can be found in [VAV+18].

6.6 Annex F: Multiple Connectivity Methods for Improving Latency Performance

1) UL scheme: If both BSs serve UL traffic, always schedule LLU first as shown in the left of Figure 6-4(a). If there is no LLU, cooperation BSs randomly schedule one of LTUs as shown in the right of Figure 6-4(b). If both users are LLU, then cooperation BSs schedule them together, and receive their packet jointly. The received packet can be decoded at *B(i)* and/or *B(j)*. BSs will apply successive interference cancellation (SIC) if the received data is not decodable. In this way, the reliability of the transmission of LLU is increased, and latency can be decreased. The receiver will try to decode the data as follows:

i) The transmission of user *i* is determined at first.

ii) If the transmission is not successful, *B(i)* tries to decode the signal from user *j*. If *B(i)* succeeds to decode the signal from user *j*, *B(i)* extracts it from aggregate interference and tries to decode the signal form user *i* again.

iii) If the signal is still not decoded, cooperation BSs try to decode the signal at the paired BS.

iv) If the signal is not decoded yet, the final attempt is SIC at the paired BS.

In this way, the proposed scheme is not necessary to decode the signal only once, but performs decoding many times by using the cooperative BS, so that the probability of successful reception is inevitably improved.

2) DL scheme: If both BSs serve DL traffic, as we explained above, always schedule LLU first as shown in the left of Figure 6-4(b). If there is no LLU, randomly schedule one of LTUs as shown in the right of Figure 6-4(b). If both users are LLU, then schedule them together, and the receiving user will apply SIC if the received data is not decodable.

i) The transmission of BS *B(i)* is determined at first.

ii) If the transmission is not successful, *i* tries to decode the signal from BS *B(j)*. If user *i* succeeds to decode the signal from *B(j)*, user *i* extracts it from aggregate interference and tries to decode the signal form *B(i)* again.

3) cross-link scheme: Lastly, if the BSs serve cross directional traffic, one BS will operate in DL (DL-BS) and the other in UL (UL-BS), or vice-versa (Figure 6-4(c)). The UL-BS uses side information sent from the DL-BS through the wired backhaul, for interference cancellation. To protect LLU, LLU is scheduled first when LLU and LTU appear together. If both users are LTU, randomly schedule one of them to reduce interference to entire network.

Due to the low mobility of BSs, we can almost perfectly know the channel state information (CSI) between the cooperation BSs. In the middle of Figure 6-4(c), *B(j)* sends the signal in the DL to user *j*, while *B(i)* receives signal from user *i* in the UL. BS *B(i)* receives interference from *B(j)*, while user *j* receives interference from user *i*. However, *B(j)* can use the high bandwidth connection to *B(i)* to send the same signal of signal to *j*, such that *B(i)* can regenerate the strong interference exploiting the data through wired stable link and perfect CSI, and cancel the interference from *B(j)* perfectly. On the other hand, the signal from user *i* remains as interference to user *j*. But, it is also possible to apply SIC to the interference from user *i* to *j*.

We consider a situation where the user transmits its data on the UL and receives its ack on the DL. If the UL transmission is not succeeded, the user cannot receive its ack and retransmits the data. The DL transmission is possibly not succeeded either. The user will retransmit the same data again. We assume that the UL time slot and the DL time slot have the same length alternately.

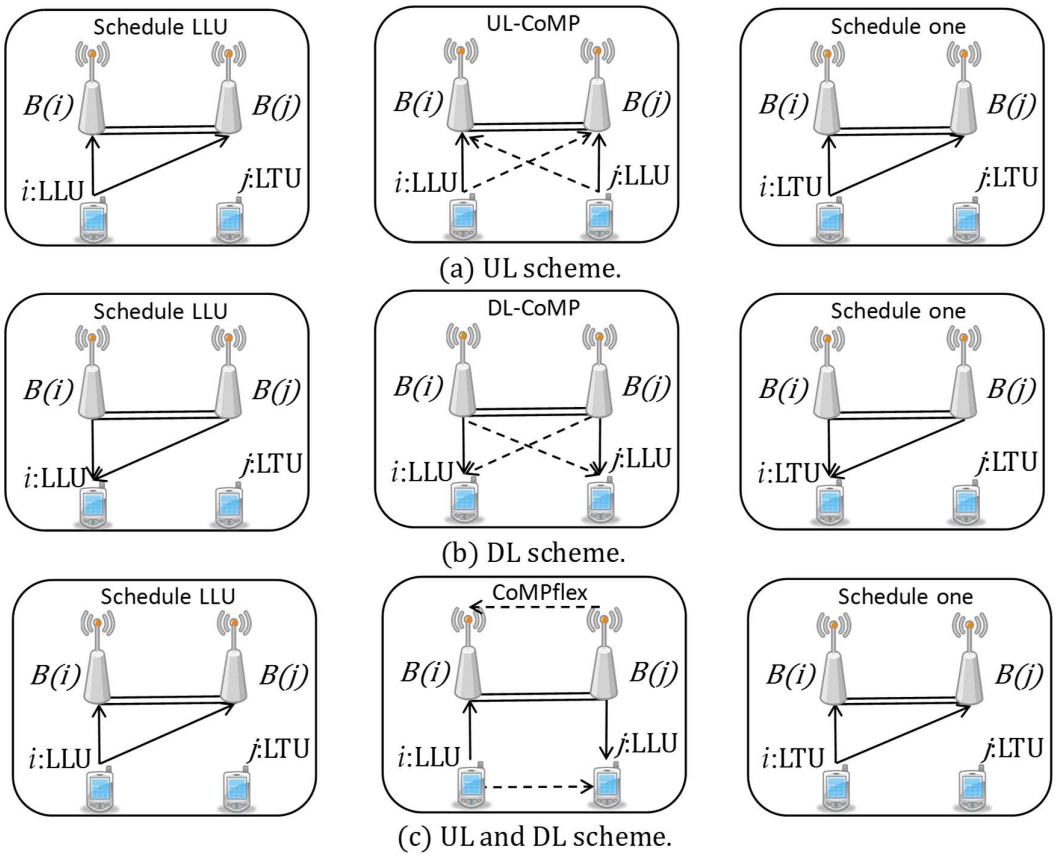


Figure 6-4: Different BS cooperation modes: (a) both BSs support UL traffic; (b) both BSs support DL traffic; (c) BSs support cross directional traffic. The users could be low-latency user (LLU) or latency-tolerant user (LTU).

6.7 Annex G: Performance Results of Low-Latency Two-Way Cellular Communications

We quantify the latency performance of control and data separation using numerical simulations. All the simulation results are obtained by performing Monte Carlo simulation with 10000 iterations. The common parameters used are shown in Table 6-2.

We validate our latency analysis with the normalized data slot length. The latency is measured by the required normalized slot length. We normalize the length of time slot as one. The two-way latency is defined as the number of time slots from when the first data was transmitted until the ack was received. The impact of the transmission success probability is shown on Figure 6-5. The two-way latency is expressed as a function of the transmission success probability. The latency is inversely proportional to the success probability of the data transmission since additional retransmissions are required if the lower is the success transmissions of the data transmission. Both Single BS and Control/Data Separation show similar trends. However, in all cases where the simulation was performed, the Control/Data Separation exhibits a lower latency than the Single BS. The difference in performance gap between the two is larger when the transmission success probability is low.

Table 6-2: Simulation parameters.

Description	Simulation Setting
Size of observation window	150 m
BS density	0.005 BS/m^2
Traffic asymmetry ratio	0.5
Noise power at MS and BS	-174 dBm
Path loss exponent	
DL SINR threshold (ACK)	-5 dB
UL SINR threshold (Data)	0 dB
BS transmission power	40 dBm
MS transmission power	20 dBm
System bandwidth	1 Hz
Simulation iterations	10000

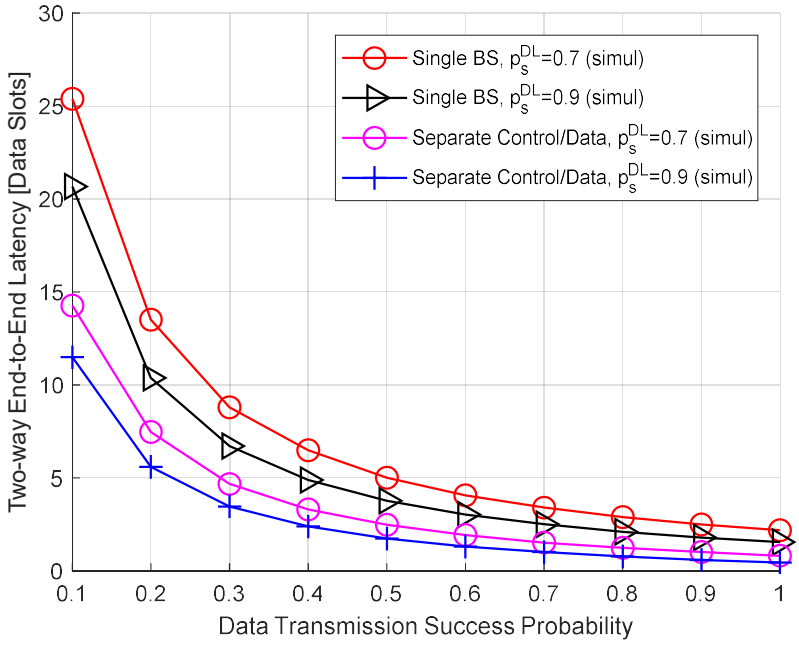


Figure 6-5: Two-way latency as a function of data transmission success probability.