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

Deliverable D6.2
5G Asynchronous Control Functions and Overall Control Plane Design

Version: v1.0

2017-04-30

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

Deliverable D6.2
5G Asynchronous Control Functions and Overall Control Plane Design

Grant Agreement Number:	671680
Project Name:	Mobile and wireless communications Enablers for the Twenty-twenty Information Society-II
Project Acronym:	METIS-II
Document Number:	METIS-II/D6.2
Document Title:	5G Asynchronous Control Functions and Overall Control Plane Design
Version:	v1.0
Delivery Date:	2017-04-30
Editor(s):	Mårten Ericson (Ericsson), Mikko Säily (Nokia Bell Labs), Panagiotis Spapis (Huawei), Shubhranshu Singh (ITRI)
Authors:	Mikko Säily, Jens Gebert, Tommi Jokela, (Nokia Bell Labs), Panagiotis Spapis, (Huawei), Mårten Ericson (Ericsson), Shubhranshu Singh (ITRI), Ji Lianghai, Nandish P. Kuruvatti (Technische Universitaet Kaiserslautern), Nico Bayer, Gerd Zimmermann (DTAG), Ingolf Karl (Intel), Alessandro Trogolo (TIM)
Keywords:	5G, Control Plane, RAN, Asynchronous functions
Status:	Final
Dissemination level:	Public

Abstract

This deliverable presents the final considerations from the METIS-II project of the 5th Generation control plane functions and the overall design. One of the characteristics of the 5G system is the ability to handle a wider range of scenarios and higher demanding service requirements than earlier systems. This deliverable presents an overall control plane framework which aims to fulfill the different scenarios and requirements.

Revision History

Revision	Date	Description
0.1	2016-11-24	Draft of Table of Content
0.2	2016-03-20	Version sent for cross-WP and PMT review
0.3	2016-03-29	Addressing comments from cross-WP and PMT review
0.4	2016-04-03	Version sent for external review
0.5	2016-04-10	Addressing external comments
1.0	2016-04-30	Final version

List of Abbreviations and Acronyms

3GPP	Third Generation Partnership Project
5G PPP	5G Private Public Partnership
AI	Air Interface
AIV	Air Interface Variants
AP	Access Point
AS	Access Stratum
BF	Beam Forming
BS	Base Station
CA	Carrier Aggregation
CAC	Central Access Controller
CMP	Control-Management Plane
CN	Core Network
CP	Control Plane
CPF	Control Plane Function
CP-H	Control Plane High
CP-L	Control Plane Low
CPM	Converged IP Messaging
CP-M	Control Plane Medium
CPNF	Control Plane Network Function
CPRI	Common Public Radio Interface
CQI	Channel Quality Indication
C-RAN	Centralized Radio Access Network
CRS	Cell-specific Reference Signal
CU	Central Unit
D2D	Device-to-device
DC	Dual Connectivity
DCH	Dedicated Channel
DCI	Downlink Control. Information
DCI	Downlink Control Information
DL	Downlink
DMRS	Demodulation Reference Signal
DU	Distributed Units
EDGE	Enhanced Data rates for Global Evolution
eLTE	enhanced Long Term Evolution
eMBB	enhanced Mobile Broadband

E-UTRA	evolved UMTS Terrestrial Radio Access
FACH	Forward Access Channel
FEC	Forward Error Correction
FFT	Fast Fourier Transformation
GPRS	General Packet Radio Service
GRE	Generic Routing Encapsulation
GSM	Global System for Mobile (communications)
GT	Group Transmission
GTP	GPRS Tunneling Protocol
GUTI	Globally Unique Temporary UE Identity
GW	Gateway
IEEE	Institute of Electrical and Electronics Engineers
HARQ	Hybrid Automatic Repeat Request
HF	High Frequency
HO	Hand-Over
HSPA	High Speed Packet data Access
ID	Identifier
iFFT	Inverse Fast Fourier Transformation
IoT	Internet of Things
IP	Internet Protocol
ITU	International Telecommunication Union
LDDE	Layer De-mapping Demodulation and Equalization
LTE	Long Term Evolution
LTE-A	Long Term Evolution Advanced
MAC	Medium Access Control
MC	Multi-Connectivity
MH	Multi-Homed
MLMDM	Modulation, Layer Mapping, and Digital Beamforming
MME	Mobility Management Entity
mMTC	Massive Machine Type communication
MPTCP	Multi-Path TCP
MRS	Modulation Reference Signal

MT	Mobile Terminated
MTC	Machine-Type Communication
NAS	Non-Access Stratum
NB	NodeB
NF	Network Function
NG	Next Generation
NGC	Next Generation Core
NIDD	Network Interface Identifier
NR	New Radio
NW	Net-Work
ORI	Open Radio equipment Interface
PCell	Primary Cell
PCH	Paging Channel.
PDCCH	Physical Downlink Control Channel
PDCP	Packet Data Convergence Protocol
PDN	Public Data Network
PF	Proportional Fairness
PRB	Physical Resource Block
ProSe	Proximity-based services
PSCell	Primary Secondary cell
QoS	Quality of Service
RA	Random Access
RACH	Random Access Channel
RAN	Radio Access Network
RAT	Radio Access Technology
RF	Radio Frequency
RLC	Radio Link Control
RRC	Radio Resource Control
RRM	Radio Resource Management
RS	Reference Signals
RT	Real-Time
RU	Radio Unit
SI	System Information
SON	Self-Organizing Network
SRS	Sounding Reference Signal
SS	Synchronization Signals
SSS	Second Synchronization Signals

S-TMSI	System Architecture Evolution - Temporary Mobile Subscriber Identity
TA	Tracking Area
TB	Transport Block
TCP	Transmission Control Protocol
TEID	Tunnel End Point Identifier
UC	Use Case
UE	User Equipment
UL	Uplink
UP	User Plane
UPF	User Plane Function
UP-H	User Plane High
UP-L	User Plane Low
UP-M	User Plane Medium
UPNF	User Plane Network Function
URLLC	Ultra-Reliable Low Latency Communication
V2X	Vehicle-to-Everything
WCDMA	Wideband Code Division Multiple Access
QCI	QoS Class Identifier
QoS	Quality of Service

Executive summary

This deliverable D6.2 presents the final considerations on the asynchronous control functions and overall Control Plane (CP) design.

The CP architecture design for 5G must be able to fulfill a wide variety of requirements such as future-proof design, high energy efficiency, higher connection reliability and tighter integration with legacy air-interfaces (AI) along with the standalone operation. Also, it is expected that 5G will operate in a wider range of frequencies (1-100 GHz) than 4G, which means that beamforming techniques may be needed to compensate for the higher propagation loss at high frequencies. The assumption in METIS-II is that the overall 5G (or NR as it is called in 3GPP) will consist of multiple different AI variants including LTE-A (AIVs), in order to handle the wide variety of requirements.

In addition to this, 5G also must handle the new service requirements set by the industry, from use cases such as extreme Mobile Broadband (MBB), Ultra-Reliable Low Latency Communication (URLLC) and massive Machine Type Communications (mMTC). The report envisions that 5G should have a common Core Network/Radio Access Network (CN/RAN) interface (denoted S1*) for both the new AIVs and the evolution of LTE-A, which enables a tighter interworking between the new AIVs for NR and LTE-A evolution improving the mobility, robustness and resource usage.

In general, it assumed that all CP functions herein are common for all AIVs regardless of the frequency band used, except possibly if network slicing is used where some specific CP functions can be turned on or off depending on the requirements. Different options on how to best split the CP or UP for different physical architecture options are further analyzed in this deliverable. The PDCP layer is assumed be used for aggregation/split layer for LTE and NR tight integration. The tight integration between LTE and NR is an important aspect of 5G already from the beginning to enable a smooth migration and higher reliability of the new 5G networks.

The control plane signalling between RAN and CN can be even further reduced by using Ethernet datagrams instead of GTP tunnels because each datagram directly contains the address information needed to forward the packet. The 5G spectrum allocation is expected to be dynamic in both time and space based on traffic demand, different type of services, as well as the different radio access technologies (e.g. LTE and NR).

One design goal for METIS-II has been to “push down” functionalities to the RAN that formerly needed CN-RAN signaling in order to reduce the CN-RAN signaling and improve efficiency. One way to do this is to enable a more RAN based paging. This concept also benefits from a new UE state between IDLE and ACTIVE, which allows the network to be able to page the UE more accurately (involving less NBs). The new state also allows for faster transition to CONNECTED mode, i.e. the UE can start transmitting faster compared to LTE due to that less required signaling.

To optimize the power consumption of mobile devices during the low activity periods while minimizing the latency for the first packet transmission from the UEs to the network, a new state called RRC Connected Inactive is proposed. The mobility and system access procedures of the new state model are configurable based on different aspects of use cases, device capability, latency access and security requirements, privacy, etc.

The initial access for 5G is an important topic in order to handle the above requirements. This report proposes several enhancements for the initial access:

- System information distribution optimized for energy efficiency, fast system access and allowing flexible deployments e.g. split of UP and system / common CP
- Coverage detection and synchronization for higher frequencies massively relying on beamforming and the low frequency layer as an access anchor layer
- Random Access procedures addressing diverse access latency requirements and for a wide frequency ranges
- Paging optimizations for RRC Connected (Inactive) UEs

Another important topic for 5G is the mobility. The report discusses the utilization of the handover procedure “make-before-break” for high demanding services to cope with the sudden signal strength drop expected to occur due to e.g. very high frequency band usage in 5G. Further on, the several new mobility concepts are evaluated including UE autonomous mobility, make-before-break and mobility concepts for URLLC. Make-Before-Break is thoroughly investigated in this deliverable, together with considerations for multi-connectivity for a centralized RAN architecture.

D2D is expected to be an integral component of 5G system and natively supported into the CP protocol stacks of 5G. This deliverable addresses several D2D components and their CP design perspective in order to enable this. One of these components is the use of D2D and context aware information, which is shown in this deliverable to be able to increase the overall capacity. Another component is the cooperative relaying, which can increase the UL user throughput.

Contents

1 Introduction

With the successful introduction of GSM, EDGE, WCDMA, HSPA and finally LTE, mobile data has now spread to almost every corner of the world. The number of mobile connections[1] is now almost 8.2 billion [GSMA+17] and still steadily increasing. On top of this, the data usage per person increases fast. The global mobile data traffic grew 74 percent in 2015 [CISCO+16]. The mobile data traffic has grown 4,000-fold over the past 10 years [CISCO+16]. Also, new challenging services such as extreme Mobile Broadband (xMBB), massive Machine Type Communications (mMTC), and ultra-reliable Machine Type Communication (uMTC) are expected to play an important role in the coming years. The extreme data growth and new challenging services are the main drivers for the development of 5G. On top of this, 5G will consist of technologies that have to fulfill ambitious requirements driven by the vertical industries[2] that are interested to use radio networks for wireless connectivity and control. A new aspect compared to previous generations of mobile systems, is the fact that the new 5G system is expected to be deployed in a wider range of frequencies, up to 100 GHz.

An important aspect of the 5G system is what type of signaling and control functions must exist to fulfill the above demands and requirements. This document is an attempt to describe these control plane aspects of a 5G system.

Note that the assumption in METIS-II is that the overall 5G AI (or NR as it is called in 3GPP) will consist of multiple different AI variants (including **evolved LTE**) – in METIS-II termed AIVs. In this document **5G**, **NR** and **AIVs** are used interchangeably. When the terms 5G, NR and AIVs are used in this deliverable it means all 5G AIVs (including evolved LTE), except where LTE is specifically specified (e.g. as in the LTE NR tight integration).

Also note that the terms for **BS** (base-station), **eNB** (enhanced NodeB), **NB**, **5G NB** and **gNB** are used interchangeably.

1.1 Objective of the document

This is the final deliverable of the overall Control Plane (CP) design for the 5G RAN design concept proposed by the METIS-II project. The objective is to report the conclusions on the design of the Radio Resource Control (RRC) protocol for the 5G RAN, including its state model and procedures associated to mobility and initial access.

This deliverable is a continuation of the deliverable D6.1 “Draft Asynchronous Control Functions and Overall Control Plane Design” [MII16-D61] but it also contains completely new content. Aspects that were covered in detail in [MII16-D61] will be summarized in this deliverable.

WP6 has two main tasks [MII-proj-spec]:

[1] Unique mobile subscriptions and M2M
[2] I.e. industries with special needs, for example mining industry, self-driving/automatic cars, mobile health-care, etc.

- Task 6.1: Design and evaluation of 5G asynchronous control functions
- Task 6.2: Architectural aspects of a common control plane framework for the 5G air interfaces

Task 6.1 is about the design of the basic control plane functions such as initial access and mobility. Task 6.2 is to investigate the architectural aspects, e.g. on investigate aspects on different protocol layer control plane split options, slicing, interworking with legacy networks etc.

Despite the focus on the RRC design, the deliverable also contains considerations on the physical layer (PHY) protocols that impact mobility and initial access, such as the design of reference signals and synchronization sequences, and assumptions related to the Core Network (CN) connectivity. It is worth mentioning that the document is not a set of protocol specifications, but rather aims at explaining, in a tutorial manner, the implications of 5G performance and system level requirements to the CP design of the 5G RAN and, in which ways the design could differ from the evolved LTE one.

1.2 Scope

The relation of this deliverable to other METIS-II deliverables is depicted in Figure 1-1 [MII17-D42] [MII17-D52].

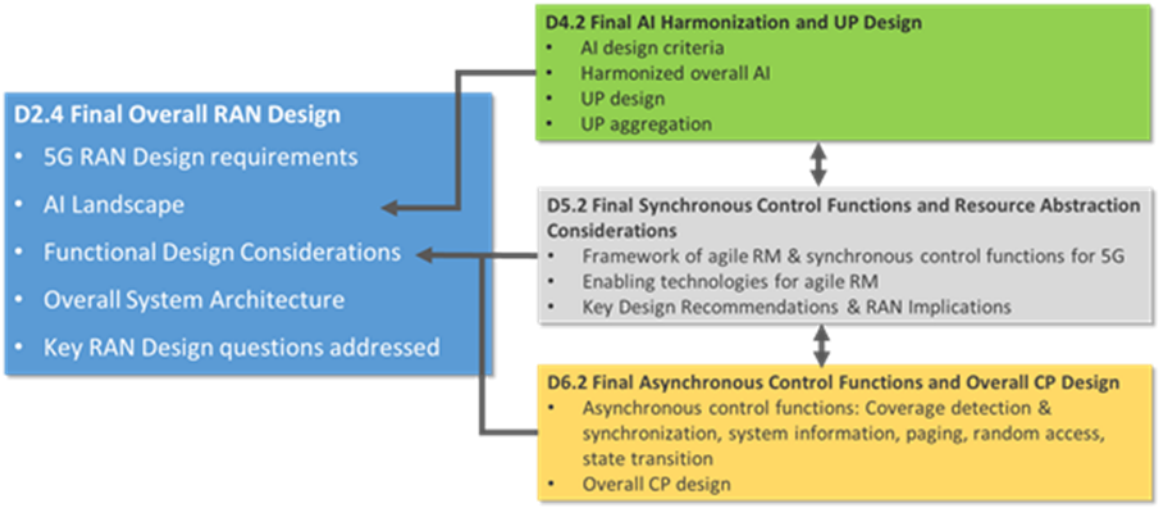


Figure 1-1 Relation of this deliverable to other deliverables.

As the successor of the METIS project, METIS-II inherits the METIS terminology to classify the CP functions into synchronous and asynchronous [MET15-D64]. The synchronous functions are the ones requiring frame/slot/sub-frame or any time-domain level synchronization between a set of functions (for instance related to scheduling, power control, etc.). On the other hand, asynchronous functions do not require frame/slot/sub-frame or any time-domain level synchronization (for instance mobility and initial access functions). While this deliverable is

responsible for the initial design of asynchronous functions and the overall CP design, [MII16-D52] describes the design of the synchronous functions.

The CP design will finally be integrated with the UP design in D2.4 [MII16-D24].

1.3 Structure of the document

The document is structured as follows:

Chapter 2 describes an overview of the **control plane framework** (which is basically this deliverable), the **control plane architecture options**, the **slicing** concept and the **SON** architecture (resource management).

Chapter 3 describes asynchronous control plane functions for Core Network and RAN control plane. The chapter describes concepts of 5G **security**, **QoS**, **Connectionless packet switching in and between the NextGen RAN and Core**, **Integration with WiFi**, and **Asynchronous spectrum** usage control. This chapter is completely new compared to [MII16-D61].

Chapter 4 describes the new **state handling** for 5G, based on lessons learnt from LTE. A solution based on a new Connected (Inactive) state is proposed. This chapter is a continuation from [MII16-D61].

Chapter 5 describes the envisioned challenges and concepts regarding **Initial Access** such as coverage discovery in a beam-based system (where coverage may be spottier), Random Access Channel (RACH) design for services with different access delay requirements and a RAN-based paging design to optimize the signaling over the CN / RAN and the radio interfaces. This chapter is a continuation from [MII16-D61] and contains several new evaluations.

Chapter 6 describes the envisioned challenges and initial concepts in the area of **Mobility.** This chapter describes the specific challenges related to the design of components necessary for the design of the mobility procedures, both for the case with active users and low active users. This chapter also presents new ideas on how to explore the usage of context awareness and data analytics to optimize the mobility algorithms. This chapter is a continuation from [MII16-D61].

Chapter 7 describes the native support of **Device-to-Device (D2D)** in 5G, including e.g. how to support efficient group based channel access, unified addressing, cooperative communication, network offloading and inter-RAT/intra-RAT Mobility Management. This chapter is a continuation from [MII16-D61] and contains a lot of new evaluations.

2 Control plane framework

2.1 Introduction

The control plane framework handles a wide range of functionalities to control the UE access to the network, such as authentication, paging, mobility, and bearer establishment. The control plane framework for 5G must be able to fulfill a wide variety of **requirements** such as:

- future-proof design,
- high energy efficiency for both the network and users,
- low latency setup,
- ability to perform network slicing,
- higher connection reliability than previous systems, as well as
- tighter integration with legacy air-interfaces (AI) along with the standalone operation.

5G is expected to operate in a wide range of frequencies and may consist of several slightly different AIVs (including evolved LTE). The aim with the CP functions/procedures is to be as **common** as possible and support all different AIV variants as well as different frequencies (including the mmW bands). There may be some exceptions of this, see Section 2.4 about slicing and RAN configuration modes. METIS-II assumes a core network / RAN split which enables an independent evolution of CN and RAN functionalities and allow multi-vendor deployments. It is further on assumed that there will be a common CN and a common CN/RAN interface (denoted S1*) for both the new AIVs and the evolution of LTE-A. This enables a tighter interworking between the new AIVs and LTE-A evolution, improving the mobility, robustness, and resource usage and minimizes the signaling to the CN. Similar enhancements are also envisioned for the evolution of the X2 interface (denoted X2* herein), which jointly with S1* become interfaces addressing multiple AIVs.

Figure 2-1 depicts some of the control plane functions which are investigated in this document. When the UE is not in active state, the control plane must support the cell selection/reselection functions/procedures (**initial access**, Chapter 5), see Figure 2-1 at number 1) and 2) box. In order for the UE to be able to connect to the system, it must listen to the system information (SI) as well as the **paging** channel (if data is network oriented). When the UE is in active state (transmitting data), the control plane must support security, **mobility** (see box 3, Chapter 6), radio bearer establishment and user **state handling** (Chapter 4) functions/ procedures. A key functionality of NR is the **multi-connectivity** (MC) ability (for higher reliability) and to handle advanced **beamforming** techniques, including BF mobility (see box 3). In addition to this, NR will support a **tight integration** with the (evolved) LTE (see box 4). **D2D** will be an integral part of 5G (see box 5, Chapter 7) and natively supported in the protocol stacks of 5G.

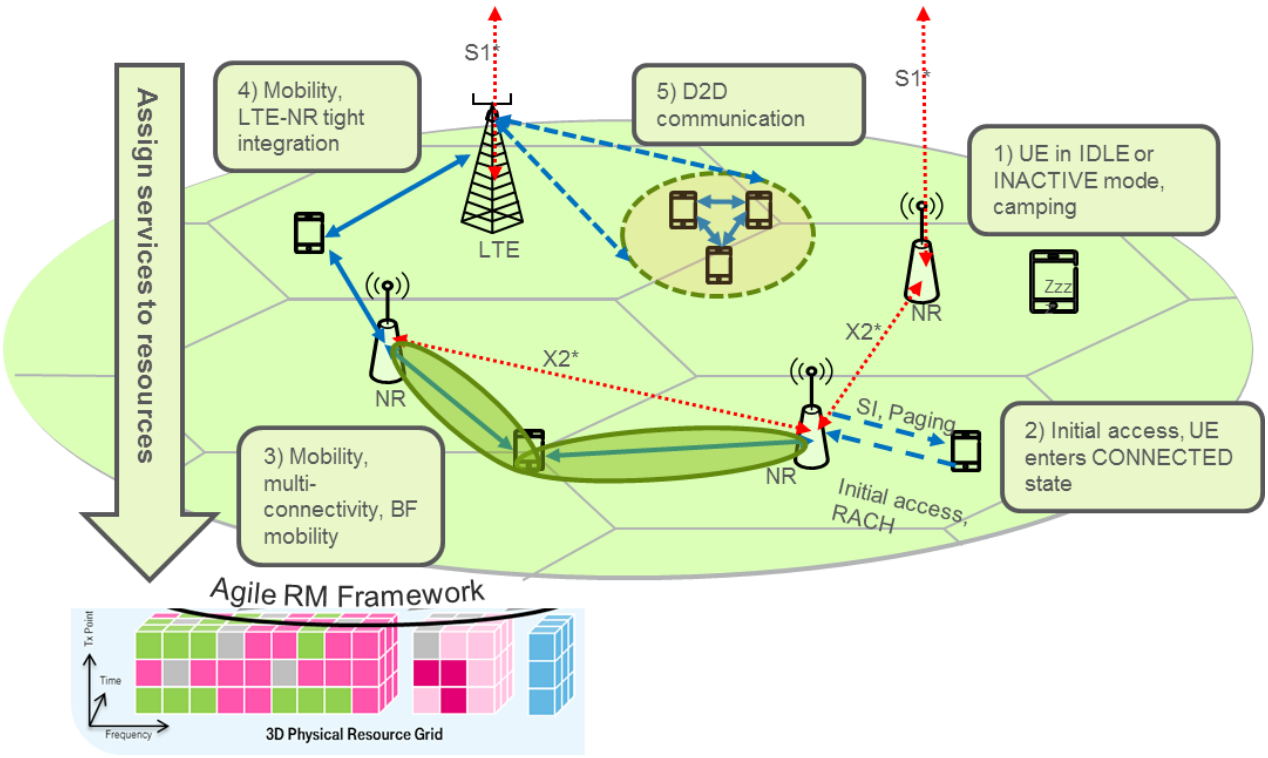


Figure 2-1: Overview of the RAN common control plane functions. Boxes 1-5 shows the asynchronous CP functions treated in this report. The Agile Resource Management is the synchronous CP functions treated in [MII16-D52].

Figure 2-1 also shows a box of the Agile RM, i.e. the so called synchronous (fast) CP functions treated in [MII16-D52]. A more thorough common CP will be presented in [MII16-D24].

2.1.1 Scope of this chapter

The remaining of this chapter deals with the **architecture** options and different impacts from different functionality. One important topic is how to split protocol layers and control plane functions in an efficient way. The fact that the proposed logical architecture has only inter-connected eNBs[3] in Figure 2-1 does not preclude architecture options with further functional splits and/or centralized/distributed options (Section 2.2 and Section 2.3). On the contrary, one of the motivations of having a simplified eNB-like architecture for the new AIVs is to enable as many functional splits as possible, bringing an additional deployment flexibility.

This chapter also investigates the architectural impacts from slicing see Section 2.4, and how to configure the different RAN slicing. It addresses which RAN functions should be common and which functions may be unique per slice, i.e. it identifies potential differences and similarities among the RAN configurations for the various services.

[3] The term for **BS** (base-station), **eNB** (enhanced NodeB), **NB**, **5G NB** and **gNB** are used interchangeably in this report.

The SON architecture and Control/Management Plane is addressed in Section 2.5 and proposes a way to improve the radio network creation, design and optimization through an End-to-End service oriented orchestration architecture through the definition of an “abstraction level” for RAN domain.

2.2 Centralization/Distribution of CP/UP network functions

This section analyses how the different functional split options can be best mapped to different deployment scenarios (physical architectures). It is important to consider how the RAN network functions (NFs) can be split over different physical entities, and which intra-RAN interfaces between the physical entities would correspondingly be needed.

Centralization of RAN NFs on the one hand provides gains in terms of centralized scheduling and flow control, etc. but on the other hand increases the x-haul (back-/mid-/fronthaul) requirements in terms of bandwidth and latency. Potential split options have already been discussed in Section 5.5.2 of Deliverable D2.2 [MII16-D22] focusing on user plane functions (UPNFs). In Chapter 4 of Deliverable D4.2 [MII17-D42] this has been extended by control plane functions (CPNFs), the interfaces between the CPNFs as well as the interface between the CPNFs and UPNFs. The resulting functional architecture addressing these so-called horizontal and vertical splits can be found in the Annex (Figure A-1).

Four different deployment options are discussed in the following: Fully decentralized CP and UP, fully centralized CP and UP, partially centralized CP and UP, fully centralized CP only. As already stated in [MII16-D22] for simplification CP/UP NFs have been structured into 3 parts with respect to their position in the radio protocol stack (-H: high; -M: medium; -L: low) with following meaning:

- CPNFs:
    - CP-H: For example, functions like AIV agnostic Slice Enabler (AaSE) [MII17-D52], High-level inter-site/air interface resource coordination like ICIC
    - CP-M: User and network specific NFs (e.g. RRC, RAN mobility, admission control)
    - CP-L: Cell configurations, Short-term scheduling, PHY layer control
- UPNFs:
    - UP-H: QoS/Slice enforcement, PDCP
    - UP-M: RLC[4], MAC, Higher PHY
    - UP-L: Lower PHY

With respect to the access network concept considering an outer and an inner layer (AN-O/AN-I) which was derived within METIS-II for the functional resource management (RM) architecture

[4] UPNF-H may also consider asynchronous functions of RLC, so only synchronous functions of RLC will remain in UPNF-M

AIV-overarching RM functionalities belonging to the AN-O layer are related to CPNFs in the CP-H and CP-M parts, whereas intra-AIV RM functionalities belonging to the AN-I layer are typically mapped to the CP-L part [MII17-D52].

Scenario 1: Fully decentralized (distributed) control and user plane (stand-alone base station)

The first option is a stand-alone base station (BS) as shown in Figure 2-2. It corresponds to a standard deployment used in 4G systems assuming a flat hierarchy of network elements. This option is characterised by having all UPNFs and CPNFs implemented in every BS (fully decentralized deployment).

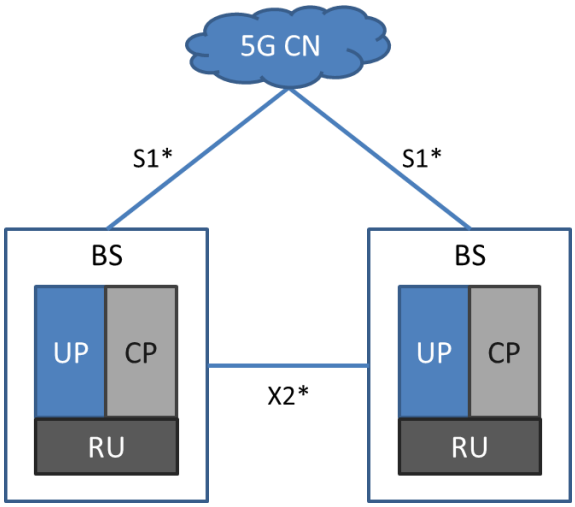


Figure 2-2: Scenario 1: Fully decentralized deployment with full RAN functionality in each BS

Scenario 2: Fully centralized control and user plane

The second deployment scenario is shown in Figure 2-3 and represents a Centralized-RAN (C-RAN) approach in which all CPNFs and UPNFs are concentrated for a certain number of radio units (RUs) at a central site, here denoted as central unit (CU). This CU is hosting the central access controller (CAC) functionality [MII17-D42] separated in the figure in CP (CAC-C) and UP (CAC-U). The split option M1 (see Figure 2-3) is the traditional CPRI or ORI split [CPRI13] [ORI14] between the CU and the so-called distributed units (DUs) which include in that considered scenario the RU only. The interface between the CU and the DUs has to carry the digital baseband data in time domain for each antenna port. Hybrid beamforming approaches as intended especially for Massive MIMO usage in mmW bands (compromise between cost and flexibility) require adjusted phase values for the analog RF precoding stage in the RU (see CP/UP interface (9) in Figure A-1 in the Annex), which necessitates extensions of the CPRI/ORI interfaces. Requirements analysis of the split option M1 can be found in D2.2 [MII16-D22] and D4.2 [MII17-D42].

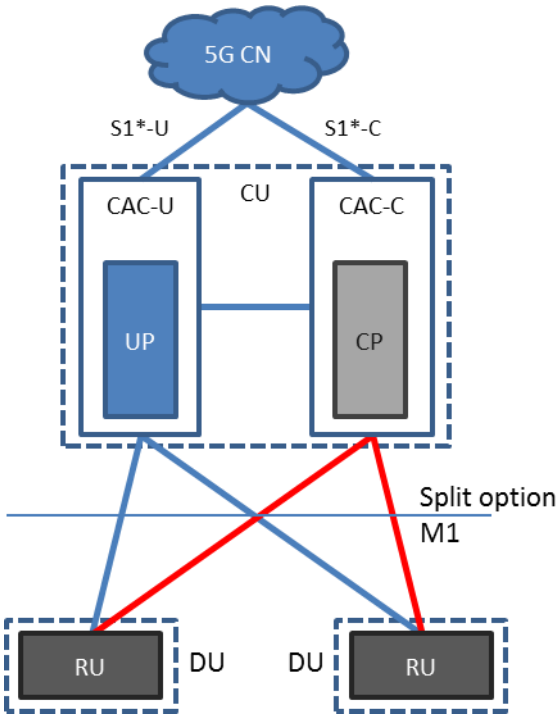


Figure 2-3: Scenario 2: Fully centralized CP/UP

Scenario 3: Partially centralized control and user plane

Two different scenarios with a partially centralized CP and UP NFs will be discussed.

Scenario 3a has a partially centralized UP and a fully centralized CP and is shown in Figure 2-4. Two different horizontal split options are suited. The M2 split requires digital baseband data per antenna port to be carried by the interface. However, the difference to M1 is that the data is in frequency domain which is less bandwidth demanding. Split M3 (see Figure A-1 in the Annex) carries user data after performing forward error correction (FEC) coding before following steps of scrambling, modulation and layer mapping/precoding. This further reduces the bandwidth requirements on the x-haul.

Requirements analysis of the split options M2-M3 can be found in D2.2 [MII16-D22] and D4.2 [MII17-D42]. A further difference between the split option M3 compared to M2 (and M1) is the fact, that is also involves additional CP/UP interfaces as shown in Figure A-1. First interface (7) between the short-term scheduler and the Modulation, Layer Mapping, and Digital Beamforming (MLMDM) block, second, interface (8) between the Cell configuration block and the MLMDM block and finally, interface (10) between short-term scheduler and the Layer Demapping, Demodulation and Equalization (LDDE) block. Due to the fact that for split option M3 the LDDE and the MLMDM blocks are implemented at the DU while the CPNFs will be implemented at the CU these interfaces have to be carried via the x-haul link. This might pose additional requirements, especially in terms of latency on the x-haul link.

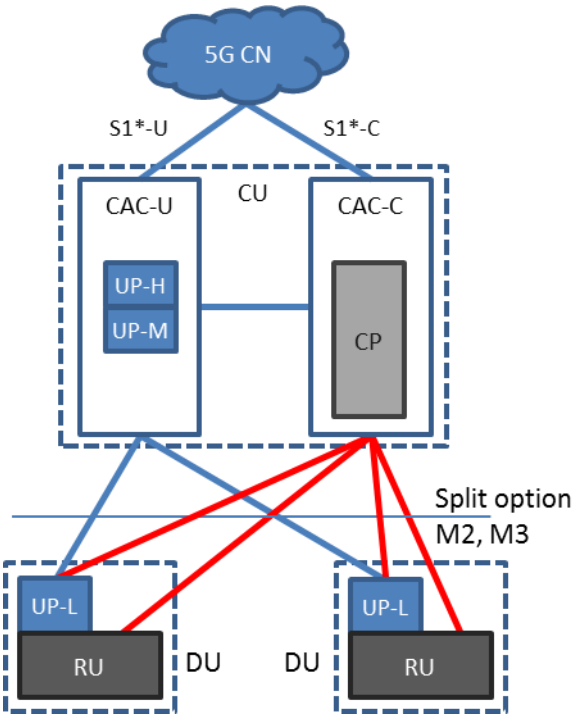


Figure 2-4: Scenario 3a: Party centralized UP with fully centralized CP

Scenario 3b is shown in Figure 2-5 in which the CAC in the CU includes only partially centralized CPNFs and UPNFs. In that case synchronous CP/UP NFs are deployed at the DUs and asynchronous CP/UP NFs at CAC-C and CAC-U, respectively. With respect to horizontal split, options M7 and M8 as shown in Figure A-1 in the Annex would fit to that approach. The difference is that with split M8 the whole RLC NF is placed at the DUs while with split M7 only the synchronous RLC part is placed at the DUs (asynchronous RLC at the CAC-U). Regarding the CP all asynchronous CPNFs stay in the CAC-C, only short-term scheduling (CPNFs-L) will be placed at DUs. The advantage of this deployment is that all CP/UP interfaces with strict timing requirements can be handled DU-internally, which also relaxes the requirements on the x-haul interface.

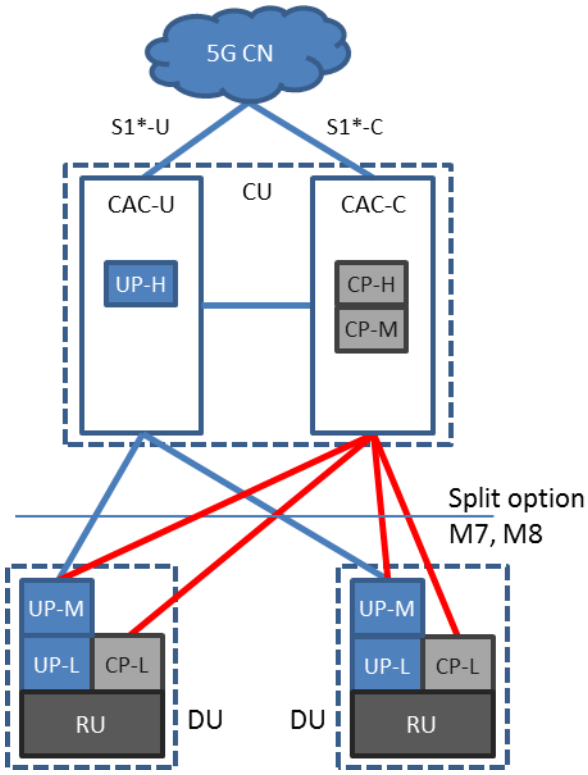


Figure 2-5: Scenario 3b: Partially centralized CP/UP allowing centralized handling of asynchronous functions for high level coordination purposes incl. multi-connectivity while still relaxing latency and data rate requirements on x-haul

As said in Section 2.1, multi-connectivity (MC) will be an important feature in 5G to achieve higher reliability than existing systems required for ultra-reliable services (e.g. for industry automation or vehicular communications). MC may be realized through radio links from collocated or non-collocated antenna sites, applying the same or even different air interface variants (AIVs) in varying frequency bands (5G NR, LTE-A Pro, WLAN, etc.) [MII17-D42] [MII16-D51] [MII16-D61]. MC can be seen as an extension of the LTE dual-connectivity (DC) approach [MII17-D42]. See Section 6 for mobility and especially Section 6.2 for the MC concept in more detail.

A multi-AIV deployment based on horizontal split option M8 in combination with the related CP/UP split, demonstrating also the needed CP split between CU and DU, is depicted in Figure 2-6. It also shows additional CPNFs hosted at the CU for e.g. quality of service and network slice control, denoted as AIV agnostic Slice Enabler (AaSE), and corresponding UP enforcement above the PDCP layer i.e., possible modification of QoS tags of incoming data flow packets if SLA and QoS policies of corresponding network slices are in danger to be violated [MII17-D52].

Due to increased opportunity range for AIV handling in a centralized environment, the ICIC CPNF is evolved to a so-called Multi-cell/-AIV Resource Mapping block which operates on an extended resource framework (antenna sites, frequency bands, AIV-related time-frequency grids, etc.). This NF also controls (via RRC) UPNFs in the PDCP layer, resulting in e.g. a duplication of data packets to be transmitted on one or more AIVs or also allowing fast switching of data streams

between AIVs in one or more DUs. Also, horizontal split option M7 can be applied if only novel 5G NR AIVs are used. For a combination of 5G NR with LTE-A Pro option, M8 has the positive aspect that it is already applied for LTE DC, thus no changes in LTE-A Pro specifications are required. Introducing option M7 also in LTE-A Pro would result in more efforts for realization and specification. More information about these higher layer CP functions can be found in Section 2.3 (from a network architecture point of view) and in Section 6.3 (with respect to inter-RAT mobility).

The approach shown in Figure 2-5 and Figure 2-6 for a high layer split strongly relaxes the x-haul requirements for 5G deployments and allows at least partial central coordination of data transmissions and receptions. The applicability is especially relevant for Massive MIMO usage, where the x-haul data rates using a lower layer split scale with the antenna numbers (ports) and therefore prevent the implementation of fully centralized CP/UP via the classical C-RAN approach.

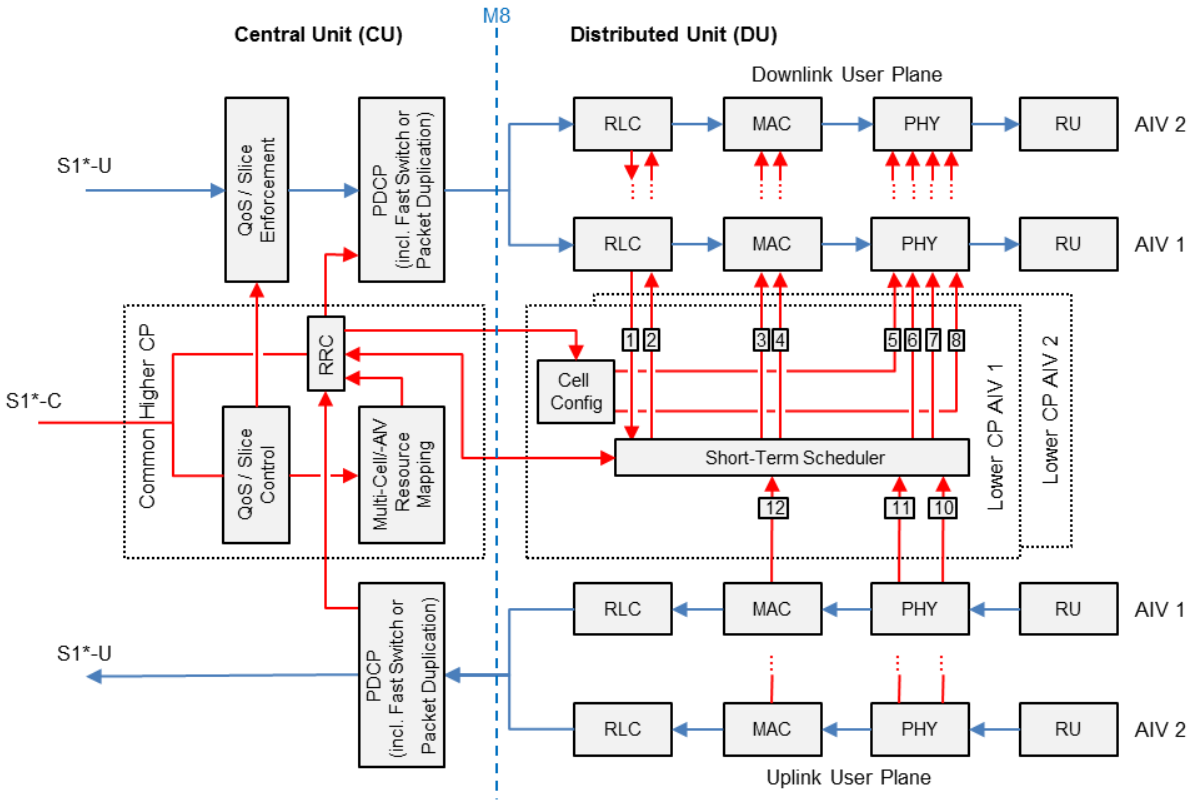


Figure 2-6: Scenario 3b with multi-connectivity: Partially centralized control and user plane deployment in multi-cell/AIV environment based on horizontal split option M8

Scenario 4: Fully centralized control plane

The last deployment scenario which would correspond to the usual software defined networking (SDN) approach with a fully centralized CP is shown in Figure 2-7. This would mean that also the short-term scheduler is implemented at the CU which poses even stricter latency requirements on the x-haul due to the fact that all CP/UP interface data, e.g. between the scheduler and the

multiplexer UPNF in the MAC layer, have to be sent via the x-haul interface. Due to the fully decentralized UP no dedicated traffic steering functions in the RAN are possible in this scenario. To implement traffic steering additional UP switching/steering functions, e.g. based on SDN approaches with CPNFs in the CAC-C, are required. Initial thoughts about those implementations can be found e.g. in [5GN17-D32].

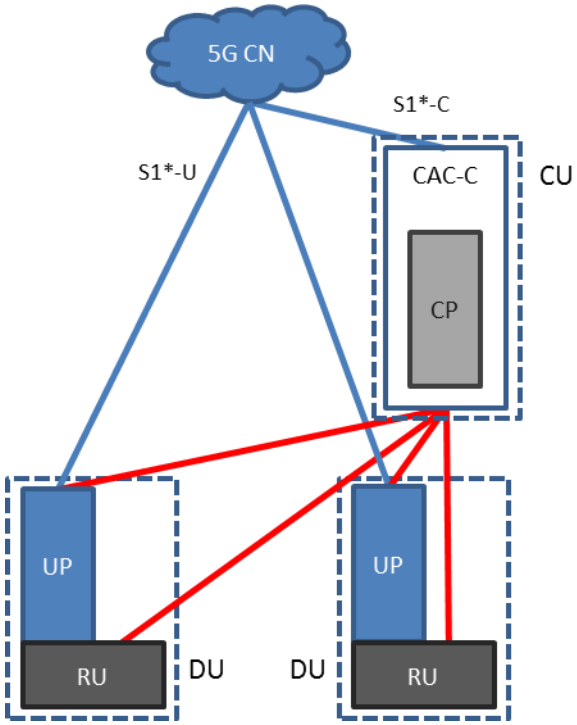


Figure 2-7: Scenario 4: Fully centralized CP and distributed UP, imposing stricter latency constraints on interfaces due to local separation of short-term scheduler and MAC multiplexer

2.3 LTE-NR tight integration architecture options

The 3rd Generation Partnership Project (3GPP) is now developing 5G, and the Dual Connectivity (DC) concept from LTE R12 is being used as a basis for a tighter integration between LTE and NR. It will enable the UE to be connected to LTE and NR (UP and CP) at the same time. Dual connectivity can increase the UE throughput due to UP aggregation (receiving data from both NBs at the same time) and make the connection more reliable. The increased reliability also comes from the case when the UE needs to switch (handover) to another Secondary eNB (SeNB), since in this case the UE can still be connected to the Master eNB (MeNB) and reliably receive RRC signaling from the MeNB. The most typical architecture for the LTE-NR tight integration will probably be the so-called bearer split “3C” options from LTE dual-connectivity, see Figure 2-8.

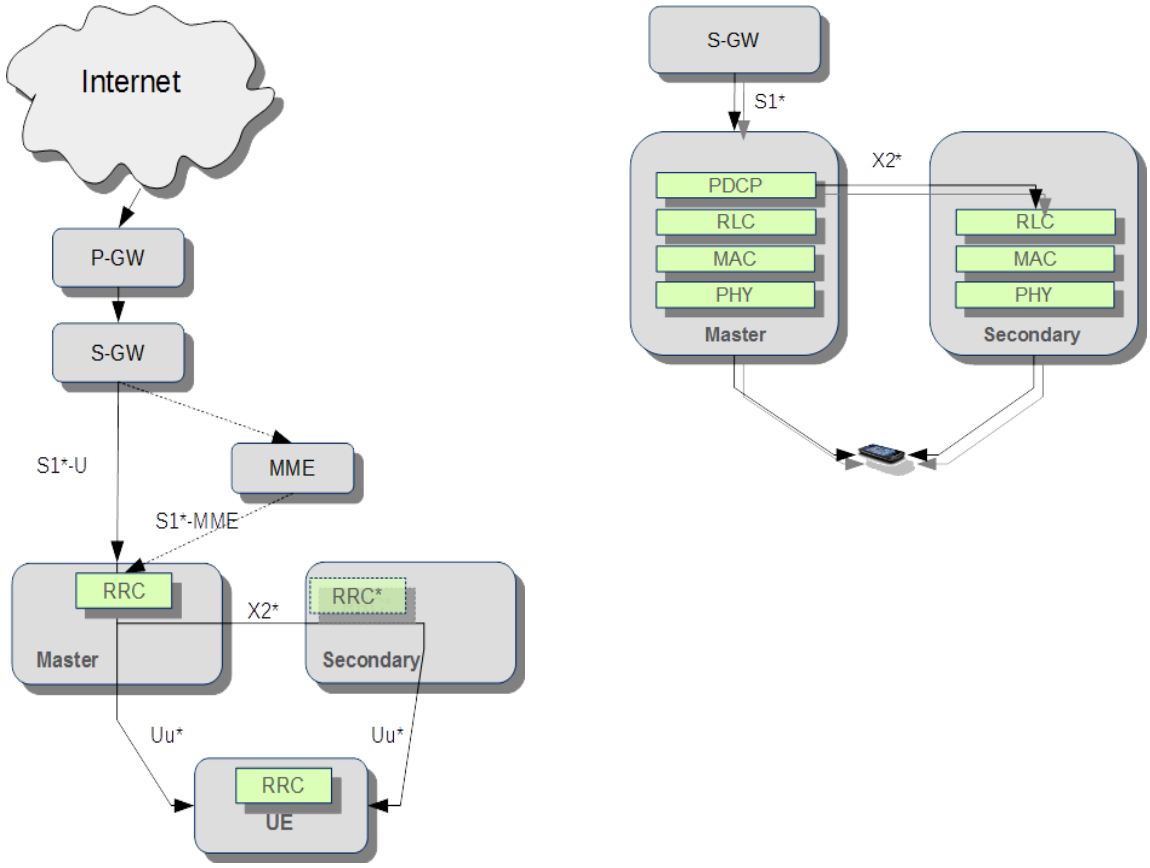


Figure 2-8 LTE-NR tight integration illustration, using the "3C" architecture alternative. Left subfigure a) shows the CP (RRC) architecture while the right subfigure b) shows the protocol stack.

In LTE DC MeNB is responsible for splitting (or aggregating in UL) the user plane data over the links. The data is sent from a MeNB lower layer to the SeNB via the X2 interface. For LTE DC, only the MeNB control plane (RRC) is connected to the CN via the Mobility Management Entity (MME). This was assumed the most likely solution in [MII16-D51] [MII16-D61] and is also the current assumption in 3GPP for 5G, i.e., a common evolved CN/RAN interface for both LTE and 5G will be used [3GPP17-38804]. This implies that no extra CN/RAN signaling is needed to add or remove a secondary node. For LTE DC all RRC messages are transmitted via the MeNB. SeNB RRC messages are sent to the MeNB over the X2 interface, and the MeNB makes the final decision of whether to transmit the RRC message to the UE. This has the advantage that there is no need for coordination, since the MeNB always makes the final decision. The disadvantage is that there is no RRC diversity and RRC messages from the SeNB take longer time since they are always routed via the MeNB. Even though it is hard to predict how the RRC for the LTE-NR tight integration will be standardized by 3GPP, it is likely that some disadvantages of the LTE DC will be addressed. That probably means that there may be duplication of RRC packets and that

the SeNB can send some RRC messages directly to UE. The transmission of RRC messages from SeNB can be achieved in fully-independent or semi-independent manner. The mobility related details of LTE-NR tight integration are discussed in Section 6.3.4.

To support different CP/UP architecture options as described in Section 2.2 some parts need to be common between LTE and NR. One option is to make the PDCP common for LTE and NR, or at least very similar, so an evolved LTE node can support (receive) NR PDCP packets and vice versa (see Figure 2-8 a). This would enable the LTE-NR tight integration to at least support Scenario 5 and 6 in Section 2.2 with rather minor standardization efforts.

2.4 5G RAN Configuration Modes for RAN slicing

There has been a lot of debate regarding whether RAN should support slicing similarly to solutions proposed for the core network or not. Although it seems straightforward to consider similar principles and solutions, we have to bear in mind that there are also significant differences between these two domains. Not all functions in RAN can meet the performance requirements if they are virtualized. These functions are operating under real time constraints especially in PHY and lower MAC layers

Using as reference the use case (UC) characteristics of [5GPPP16] which builds on the METIS inputs and combines the UC characteristics of various 5G PPP projects (i.e., METIS-II, NORMA, CHARISMA, FANTASTIC5G, mmMAGIC, SPEED-5G, etc.) we can identify the key characteristics of each UC and the respective ranges. In particular, the following key characteristics of each UC have been extracted:

- Device density
- Mobility
- Traffic Type
- Data Rate
- Latency
- Reliability
- Availability (coverage)
- Traffic applicability

The aforementioned categories should be analyzed for identifying which of the aspects of each function will be affected by the extensive diversity in the UCs. Please note that the last category included has not been identified in the analysis of the UCs from the projects but it has been included in this analysis since special differentiation is required if we refer to local traffic rather than traffic coming from wide area sources (e.g., internet). The actual effect of each characteristic that it may have to a function is presented in Table 2-1. Please note that Table 2-1 contains two parts, one for the UP and one for the CP. What emerges from the previous analysis, only few use cases families (i.e., integrating many UCs in each family) is expected to be supported by RAN.

Table 2-1: RAN functions affected by the UC characteristics

UC Characteristic	Affected Function for User Plane	Affected Function for Control Plane
Device density	Waveform, coding, access, scheduling	Waveform, coding, access, scheduling, resource allocation/management, interference management
Mobility	PDCP, RLC	Location Management, Handover
Traffic Type	resource allocation, scheduling, access (connectionless case)	Session Management, resource allocation, scheduling, access (connectionless)
Data Rate	PHY numerology, scheduling	PHY numerology, scheduling, Session management (dual connectivity)
Latency	PHY numerology, scheduling, access (connectionless), RRC	PHY numerology, scheduling, access (connectionless), RRC
Reliability	Waveform, coding, HARQ, RLC	Waveform, coding, HARQ, RLC, RRC (dual connectivity, multi-hop)
Availability (coverage)	scheduling, power control, interference management	scheduling, power control, interference management
Traffic applicability	Routing/forwarding protocols	Session management, handover, mobility management

A limited number of combinations of different RAN functions can serve the described UCs. This can be justified by the UC requirements and the respective grouping, as well as the more restricted ability of virtualization in the RAN. Each combination may be called RAN configuration mode (RCM) which is a composition of RAN network functions, specific function settings and associated resources (HW/SW, and network resources). An RCM can be statically defined or fully flexible, and this is up to the implementation and the requirements for flexibility and future-proofness (i.e., in case a totally new UC arises with new unforeseen requirements). The decision whether the RCM is statically or fully flexible relates to the UCs that will be covered by the 5G system. It can be foreseen that in the earlier phases of the 5G deployment the first option can be followed but moving towards the new and more ambitious UCs (such as URLLC, or specific V2X use cases) the fully flexible RCM could be more suitable.

In any case the number of the different RCMs will be linked to a large number of core network slices that will cover various businesses. In other words, an end to end slice will comprise a PHY layer configuration, together with certain tailor made RAN configuration, and a core network slice.

Note that our analysis is consistent with the current understanding in 3GPP where the current agreement is that RAN should be slice aware via some explicit or implicit identification (e.g., based on an abstraction model) [SMA+16]. This implies that the RAN should be able to differentiate at some degree at least the different slices and treat them differently. This will facilitate the flexible use of resources among slices in the RAN.

Figure 2-9 depicts three exemplary RCMs. In particular, it can be seen that the different RCMs share an RRM (Radio Resource Management) function for ensuring the sharing of the common radio resources; also, this function can ensure that, in the case of the RCMs sharing the lower layer functions the slice isolation can be ensured at least using QoS classes. The inter-slice RRM can further ensure the meeting of the requirements of each UC by considering the available resources and by using the proper RRM technique (e.g., interference coordination, muting, etc.). Each slice anyway can apply its own RRM strategies according the slice specific characteristics.

For the RRC part, as it is seen there is a shared part which enables the slice selection – alternatively this can be achieved by a common slice which will provide information for slice selection. Each slice can have its own RRC functions and configurations as well so as to tackle the special UC requirements when it comes to particular functions (e.g., DRX, DTX, measurements reporting, TAU periodicity, cell selection strategies, etc.) when particular shavings can be achieved.

For lower layers such as the PDCP and the RLC, depending on the message size, or the delay requirements certain functions can be either omitted (e.g., header compression, ciphering) or modified (e.g., segmentation, re-ordering, ciphering). As it may be seen in Figure 2-9, even more drastic approaches can be used for certain RCM, where a combination of the functionalities of two protocols can take place (RLC and PDCP in the uMTC RCM). The RCMs that share the lower layers (PHY, MAC, etc.) should have a joint "Unified Scheduler" for enabling them to share the resources more dynamically. On the other hand, other RCMs have their own separate resources and functions based on their key characteristics. These functions may be unique for an RCM or the same for more than one RCM but configured differently. Finally, for some physical layer functions these can also be configured to the slice requirements,

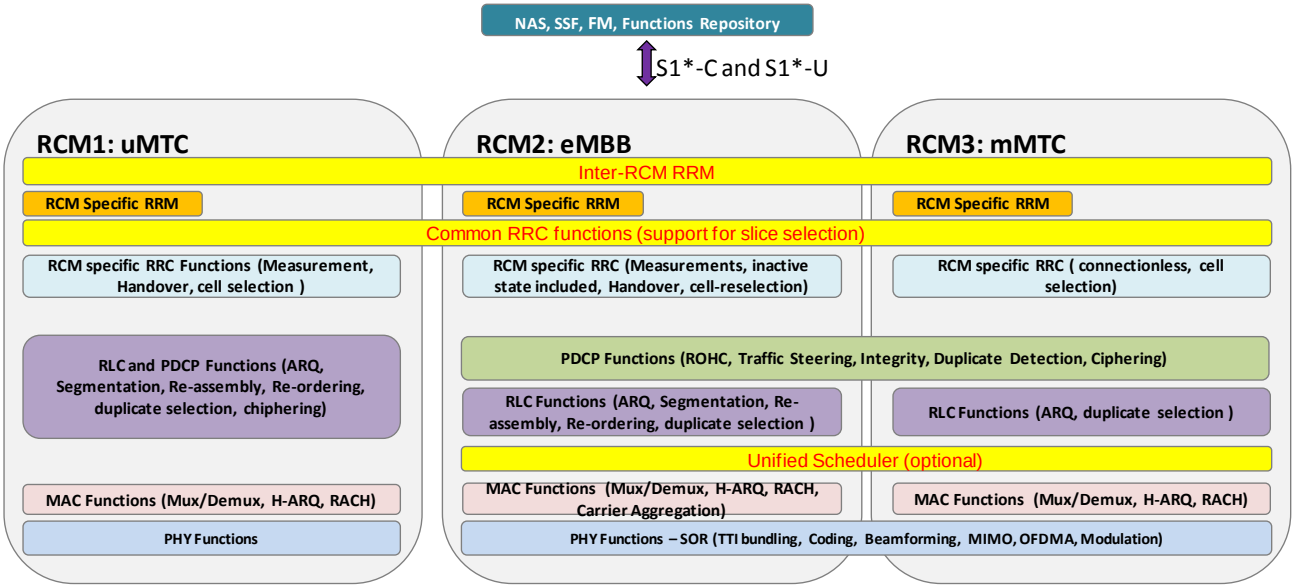


Figure 2-9: Example of RCMs with shared and independent functions

2.5 SON architecture and Control-Management Plane

Recent years have seen a rapid development of mobile access networks with an increasing complexity due to the following factors:

- Deployment of a plurality of radio technologies in many frequency ranges, some of which are shared between different technologies

- Wide range of user equipment operating in the network, with different capabilities and features
- Network nodes with different vendor implementations
- Operational requirements and constraints for the full respect of the regulatory framework

With the introduction of 5G, it is expected that such complexity will increase further, requiring the introduction of new and more flexible solutions. In such a context, SON will play an essential role by the realization of an evolved "closed loop" management, collecting network elements quality indicators and, based on these, by identifying in "real time" the actions to improve the performance and/or solve local problems in the network. With the advent of 4G, the mobile access evolution gave a first response to the increasing complexity of network management, by introducing features “distributed” in network elements able to automatize some basic configuration and optimization processes (e.g. related to mobility support, traffic balancing, improvement of the radio signal coverage, etc.). The main feature of “distributed SON” approach (D-SON) is the capability of fast closed loop reaction to network events. In addition to D-SON paradigm, a “centralized SON” approach (C-SON) has been also introduced in the network management domain, in order to enable a unified and coordinated closed loop control over multiple network nodes, considering also different radio access network technologies, while allowing less stringent reaction timing. If on one side the C-SON approach is an enabler for integration between RAN features (including D-SON) and mobile network operators’ applications, on the other side it may suffers of a lower flexibility in multivendor scenarios.

To overcome such limits, a possible solution can be an “Open SON” architecture (Figure 2-10) based on the following pillars:

- Wide availability of open interfaces (i.e. Application Programming Interfaces, API) for an effective data/command exchange between the functional blocks of the architecture. Such APIs would enable the communication with the NDO (Network Development Optimization) tools of the mobile network operator and the EM (Element Manager) functions, in conjunction with the standardized management interface Itf-N [3GPP09-32101].
- Various levels of programmability, in order to assure the needed flexibility for managing a constantly evolving radio access network
- Interworking with the tools developed by the mobile network operator for the radio access design and optimization

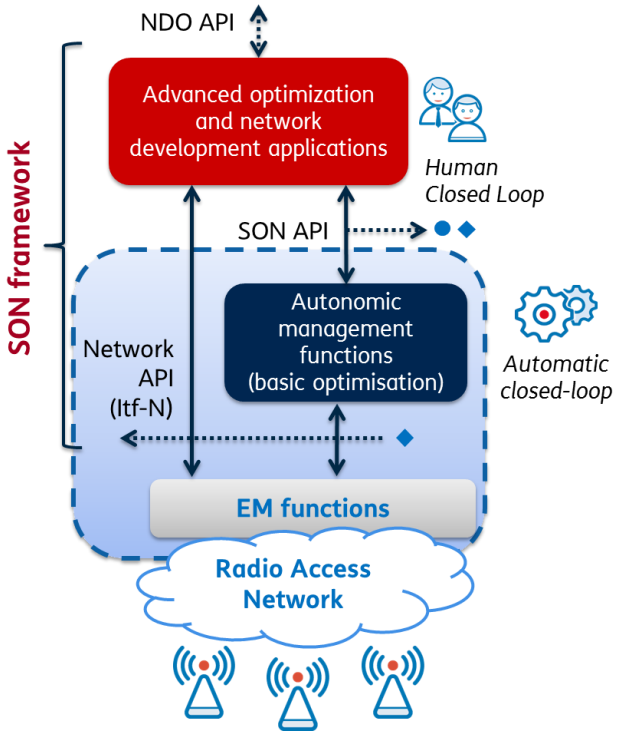


Figure 2-10: "Open SON" architecture

As depicted in Figure 2-10, the "Open SON" architectural approach enables two different "closed loop" optimization processes:

- "Automatic Closed Loop" addressing basic optimization and configuration activities, for which a complete automation can be envisaged
- "Human Closed Loop" addressing more complex activities, where the radio access specialists are supported by software applications, enabling also an effective control of automatic functions (through APIs)

Such activities are defined in order to reach two different targets:

- Improving radio network creation, design and optimization processes, by introducing automation and flexibility
- Defining an "abstraction level" for radio access network domain, enabling the integration in End-to-End service oriented orchestration architecture

The "Open SON" architecture has been conceived to manage in a more efficient and integrated way the current radio access networks. Thanks to its abstraction and flexibility, the "Open SON" architecture supports also the evolution towards the so called "Virtual RAN", based on the NFV (Network Function Virtualization) paradigm within the radio access domain. Current radio access network generation will evolve and will be complemented by new radio technologies, improving radio performance and enabling new services: bit rate up to several Gbit/s to allow multimedia

contents, end-to-end latencies of few milliseconds to allow real-time services, high reliability for mission-critical and public-safety applications, and massive connection of IoT devices. In order to efficiently and flexibly support such a plurality of use cases, 5G access architecture is expected to be designed in order to allow different deployment options (e.g. flexible functional split introduced in Section 2.2) and to aggregate/coordinate network resources adaptively by considering specific service needs (see Section 2.4). The “Open SON” architecture has to evolve in this context, in order to manage new generation technologies and to extend the network management model, enabling an effective network slicing in the radio access domain.

In the context of LTE networks, the communication between the distributed SON functionalities in the network nodes and management has been specified as extensions of the pre-existing interfaces of the related protocols, therefore divided for Control Plane (X2 and S1) and Management Plane (Itf-N). This architecture facilitates the interoperability for the SON functionalities between the network nodes (over the Control Plane), something that is more difficult to implement between the distributed functionalities and the centralized ones (Management Plane) that traditionally are implemented in the context of the same network vendor [NGMN14]. Aiming to an efficient management of the high dynamicity and heterogeneity of a 5G network, is possible to conceive the introduction of a dedicated Control-Management Plane (CMP), provided of an own specific communication protocol (CMP protocol) and able to interact with the AIV specific control and user planes (Figure 2-11).

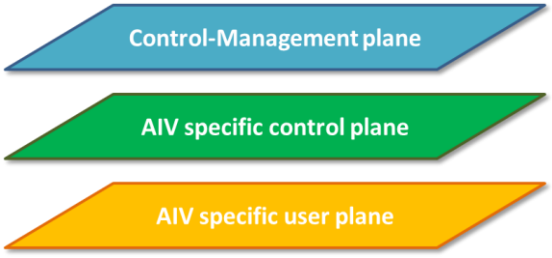


Figure 2-11: Control-Management Plane with respect to the traditional control/user planes

The CMP protocol would enable the realization of autonomic functionalities in a 5G network (dynamic and heterogeneous), bringing together management functionalities (configuration, performance, fault, trace) in a manager/agent (or client/server) manner and peer-to-peer control functionalities. In particular, in a single framework it would be possible to instantiate centralized or distributed SON functionalities and relocate them from a centralized cloud-based environment to distributed physical nodes of the access network as well as create dynamic relationships between the functionalities based on their physical location. In addition, such approach would enable to decouple the logical SON framework from the radio technologies (AIV) and from the respective network architectures.

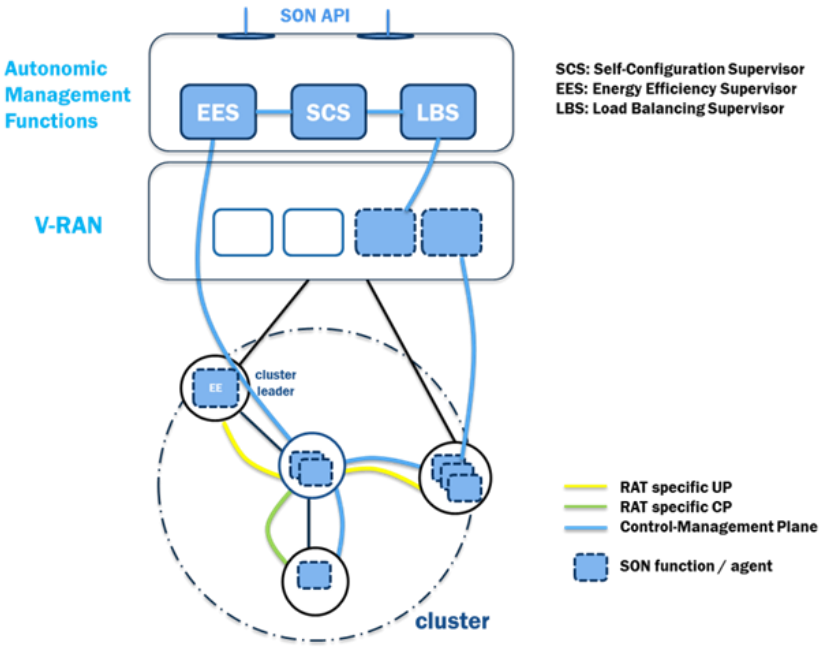


Figure 2-12: Control-Management Plane functional architecture

In Figure 2-12, the supervisor functionalities such as Self-Configuration Supervisor (SCS), Energy Efficiency Supervisor (EES) and Load Balancing Supervisor (LBS) are represented as centralized cloud functionalities that use the data provided by the distributed nodes through the related agents, thus realizing different SON processes such as, for example, self-establishment, self-adaptation, energy saving management, load balancing, etc. Nevertheless, it is also possible to position the SON functionalities locally (for example by instantiating an EE autonomic function at a cluster level). The communication between the supervisor functionalities and the autonomic functions/agents is achieved through the CMP protocol that can use the user plane (yellow lines in Figure 2-12) or the control plane (green lines in Figure 2-12) of the different supported AIVs. In relation to the configuration actions, the CMP protocol foresees to manage the peer-to-peer relations between the various SON supervisors and the SCS (Self-Configuration Supervisor) in charge of the network configuration supervision.

In Annex A.9, some examples of CMP protocol use cases for 5G deployment scenarios are analyzed. For each use case is also presented a possible implementation described through the CMP protocol defined previously. As the current classification, the analyzed use cases are subdivided in *self-configuration* (related to the autonomic functionalities for the configuration of the network) and *self-optimization* (related to the autonomic functionalities for the network optimization). Alternative solutions that act on a shorter time frame are investigated in the METIS II deliverable D5.2 [MII17-D52].

3 Core Network and RAN Control Plane interaction

Next generation wireless networks provide communication capabilities for new use cases and business as well as a myriad of new devices, applications and services. Therefore, next generation wireless networks drive many new requirements for all protocol layers and functions including the asynchronous control plane ones. This chapter deals with the interaction between the CN and the RAN, e.g. signaling and functionality spanning both RAN and CN. It discloses exemplified first asynchronous control plane concepts for 5G security, QoS, connectionless packet switching in and between the NextGen RAN and Core, the integration with WiFi, and the asynchronous spectrum usage control.

The RAN 5G security Section 3.1 concludes that the new Connected Inactive state needs new security solutions. Also, it is also concluded that massive MTC services need special security solutions to work efficiently. Section 3.2 describes a concept to reduce the control plane signaling between RAN and Core by using Ethernet datagrams instead of GTP tunnels because each datagram directly contains the address information needed to forward the packet. Section 3.3 briefly discusses different options for 5G QoS. Section 3.4 addresses the so called multihoming aspects of 5G, i.e. when a 5G UE is connected to another network simultaneously: When WiFi access points are co-located with 3GPP sites, tight integration provides large gains. If not, network assistance by sending average load information of each system to UEs is a good solution compared to a blind upper layer integration. Section 3.5 deals with efficient use of the spectrum considering the dynamic (in time and space) traffic demand, the different type of services, as well as the different radio access technologies (e.g. LTE and NR)

3.1 Security

Security prevents eavesdropping and manipulating of the data from third parties and therefore security is of vital importance for 5G, as for previous systems. With the increasing importance of the 3GPP mobile system in today's society security is more and more important. Since the computational abilities to perform a security breach increases there may be a need to update and rethink the security procedures of previous wireless network generations for 5G.

A new challenging area for security improvements is for example the Massive MTC scenario. Assuming the massive number of devices for 5G systems, the distribution of security keys across network nodes may affect the latency negatively. This section focuses on the parts related to the 5G RAN and mobility in particular.

3.1.1 State of the art

The security for LTE includes the following domains:

- Network access

- Network domain
- User domain
- Application domain

In this METIS-II contribution we handle the **network access security**.

The network access domain in LTE comprises security features for authentication of a user to EPC. This is including **mutual authentication**. Mutual authentication means that both the UE and the NW must acknowledge themselves, i.e. they must have the knowledge of the secret key (in principle the user's IMSI number). Both the control plane and the user plane are protected with scrambling (**ciphering**) of the data using cryptographic keys. These keys are derived based on the user's IMSI number. A key aspect in LTE is to use **key separation**. This means that different keys are used for authentication, **authorization** (check the user subscription etc.), ciphering of CP and UP. Also, when a new access is done or a new cell is accessed by the same user, new keys are derived.

[3GPP17-38913] Section 10.11 specifies first security requirements for next gen systems:

> *"The RAN design for the Next Generation Radio Access Technologies shall ensure support for integrity and confidentiality protection of radio signaling messages, including messages between RAN and Core network nodes.*
>
> *The RAN design for the Next Generation Radio Access Technologies shall ensure the ability to support integrity and confidentiality protection of user plane messages, including messages between RAN and Core network nodes, with the use of such security to be configurable during security set-up.*

The requirements in [3GPP17-38913] Section 10.11 show that LTE security can support most of the requirements in terms of integrity/confidentially protection and user privacy. However, there are some areas that need to be improved for 5G in order to fulfill the requirements for 5G above [3GPP17-38913].

- Mobility enhancements for security to support the targeting zero ms interruption time
- Inactive state transmission, a solution to further improve UE state transitions down to below 10 ms between inactive and active state
- Tight interworking, support for tight RAN level inter-working between NR and LTE

3.1.2 Security in 5G RAN

The following areas are identified in METIS-II and options for a way forward are given for updated or new security solutions for 5G RAN.

Mobility

As said above, in LTE a handover always implies derivation of new keys and this shall also be possible for 5G. However, for services requiring zero ms delay it is challenging to reach zero ms interruption and at the same time change the keys. One way to solve this is to use multiple keys in parallel, or that the same key is used for transmission/reception in both base stations (meaning that the moving of security context is decoupled from the radio switch).

Connected Inactive state transmission and connected Inactive mobility

The option to transmit small data without leaving Connected Inactive was discussed in METIS-II at Stockholm meeting starting on Sept 2015, e.g. before the 3GPP NR SID started. This was effectively the discussion of possibility to start UL and/or DL data transmission in Connected Inactive without state transition to Connected Active mode. There was noted that security support is needed.

One motivation for the new Inactive state in NR proposed by METIS-II was that the UE shall be able to faster start a new data transmission. In LTE each new session required a new key. Thus, the security solution must allow that the UE can start UL data transmission at an early stage. The UE must therefore receive the necessary keys before the UL transmission.

Whenever non-scheduled UL is available at UE for sending, UE uses the associated bearer for the data. The user data is ciphered in PDCP layer and the UE ID and shortMAC-I are appended to the UL data for identification and integrity protection respectively

UE based mobility during Connected Inactive state needs to be updated with new security when the UE makes a cell reselection to a new gNB and initiate access. The reason for this is that since the access is not only about the key derivation but the overall Context including the DRBs and SRBs. It is likely that a Context fetch is needed.

Support for tight RAN level inter-working

METIS-II is a strong supporter of a tight integration between LTE and NR. This was also agreed by 3GPP. The tight integration mobility is controlled by the RAN and to support this it is beneficial if the security context retrieved from the 5G CN is the same or harmonized for LTE and NR. Thus, the UE establishes the security in one RAT, which is then also used in the other RAT (e.g. to derive keys to be used in the other RAT). Another way to solve it is to use multiple keys in parallel for LTE and NR.

An application might be the support of the expected massive number of MTC devices, where the distribution of security keys across network nodes must be made in an efficient and fast way, so the latency is not sacrificed. One option to do this is to use the originating key for many MTCs in the same area but slightly modified for each MTC device. The originating key can for example be the same key derived for a normal connection. The key per MTC device is then derived based on that. The time to derive the key for each MTC should be faster than deriving the originating key. The originating key can be changed on the demand basis by the eNB.

3.2 Connectionless packet switching in and between the NextGen RAN and Core

3.2.1 Scope of connectionless packet switching

This section presents a connectionless packet switching solution for the interface between the RAN and the Core network and other RAN/Core internal interfaces. The solution is based on the Ethernet protocol and uses locally administered IEEE MAC addresses for the 3GPP UEs. In the current 4G system, the user plane on the RAN-Core interface (S1-interface), inside the 4G RAN (X2-interface) and inside the Core network (S5, S8, S2a interfaces) is based on the GPRS-Tunnelling protocol (GTP). Connection management procedures like Connection setup, modification and release procedure are required which exchange connection identifiers like GTP Tunnel Endpoint Identifiers (TEIDs) via the control plane before the user plane can be used, e.g. when a UE starts a new session or enters the active mode.

The advantage of connectionless packet switching is that this exchange of connection identifiers like the TEIDs is not necessary because each user plane packet directly contains the address information needed to forward the packet. Therefore, the amount of control plane signaling can be reduced. The advantage of the reduced amount of control plane signaling is especially important in cases where a UE transmits only sporadically small amounts of data (mMTC use case) and where the state-of-the-art tunnel setup adds a large control plane signaling overhead which is reduced with the connectionless packet switching approach.

Connectionless packet switching provides an instantaneous data delivery which is realized by using the full address information in the header of each packet.

3.2.2 Ethernet with locally administered IEEE MAC Addresses

In the connectionless packet switching approach, a connectionless protocol like Ethernet is used on the user plane between the RAN and the Core (S1*-Interface, also known as NG3 interface) as well as inside the CN (NG9 interface). The 5G Network Architecture with an Ethernet based User Plane is shown in Figure 3-1.

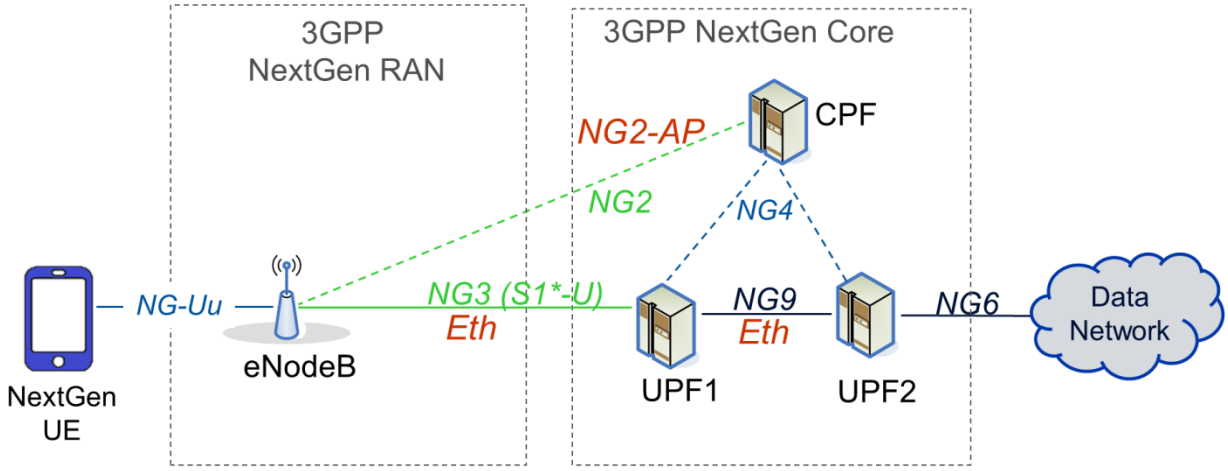


Figure 3-1: Ethernet based User Plane Protocol Stack.

In the connectionless packet switching approach, all relevant information like end point identifiers, QoS class identifiers or end-to-end protocol identifiers are included in the Ethernet protocol header which is shown in Figure 3-2. This is different to the tunnelling procedures using either GTP-tunnels or GRE tunnels with a key extension used as tunnel identifier. Instead of negotiating tunnel parameters with a control plane protocol, all control information is included in the user plane header.

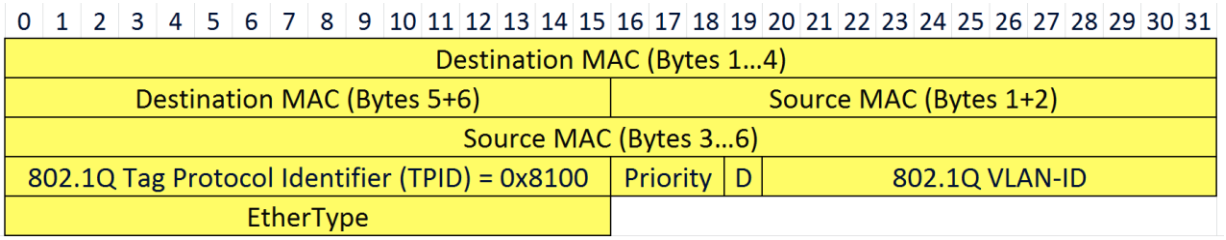


Figure 3-2: Ethernet protocol header

In order to enable this Ethernet based solution, the 3GPP network interface(s) of a UE need to be identified with an IEEE MAC Address. So far, 3GPP UEs do not have an IEEE MAC Address. As solution, we propose to use a locally administered IEEE MAC address for the UEs network interface which is created from 3GPP identifiers identifying the UE (UEID) and a Network Interface Identifier (NIID) identifying the 3GPP network interfaces of that UE. This mapping is shown in the following equation:

$$MAC_{UE,NI} = f(UEID, NIID)$$

As UEID, an existing identifier from 3GPP shall be used. The International Mobile Subscriber Identity (IMSI) however should not be mapped into an IEEE MAC address for security reasons. The Globally Unique Temporary UE Identity (GUTI) with a length of 80 bits is too long in order to be mapped into a 48-bit IEEE MAC Address. The SAE Temporary Mobile Subscriber Identity (S-

TMSI), which consists of the lower 40 bits of the GUTI, locally identifies a UE within a MME group. With the length of 40 bits, the S-TMSI can easily be mapped into a 48-bit IEEE MAC Address. Therefore, we propose to use the S-TMSI as UEID [GZ16].

As NIID, one option is to use the 3GPP ServCellIndex as defined in the 3GPP RRC Protocol Specification [3GPP17-36331]. The ServCellIndex is a short identity used to identify a serving cell (i.e. the Primary Cell or a Secondary Cell). Value 0 applies for the Primary Cell (PCell), while the SCellIndex as also defined in [3GPP17-36331] applies for Secondary Cells. The ServCellIndex is currently defined as INTEGER (0…31)) and thus requires 5 bits.

The mapping of the UEID (S-TMSI) and the NIID (ServCellIndex) into a locally administered IEEE MAC address is shown in Figure 3-3.

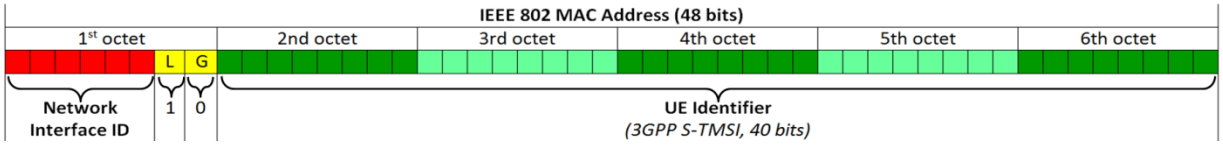


Figure 3-3: Locally administered IEEE MAC address including 3GPP identifiers

3.2.3 Evaluation

The Attach procedure using the connectionless packet switching approach with Ethernet as user plane is shown in Figure 3-4. The procedure starts with an Attach Request sent from the UE towards the eNodeB, which is then forwarded to a Control Plane Function (CPF) performing similar procedures as an MME. With message 2 *Initial Context Setup Request/Attach Accept*, the eNodeB receives the actual GUTI, which includes the S-TMSI. From this information, a UE specific MAC address can be created as described in the previous section and after the data radio bearer establishment on the air interface, user plane data can be sent in the uplink. The eNodeB uses the UE-specific MAC address as Source address and the MAC address of the User Plane Function (UPF) acting as PDN-GW as destination address. The UPFs learn the MAC addresses of a UE from these uplink messages and are then also able to send downlink user plane data towards the UE. The attach procedure is finalized with the messages 3 *Attach Complete* and 4 *Initial Context Setup Complete*.

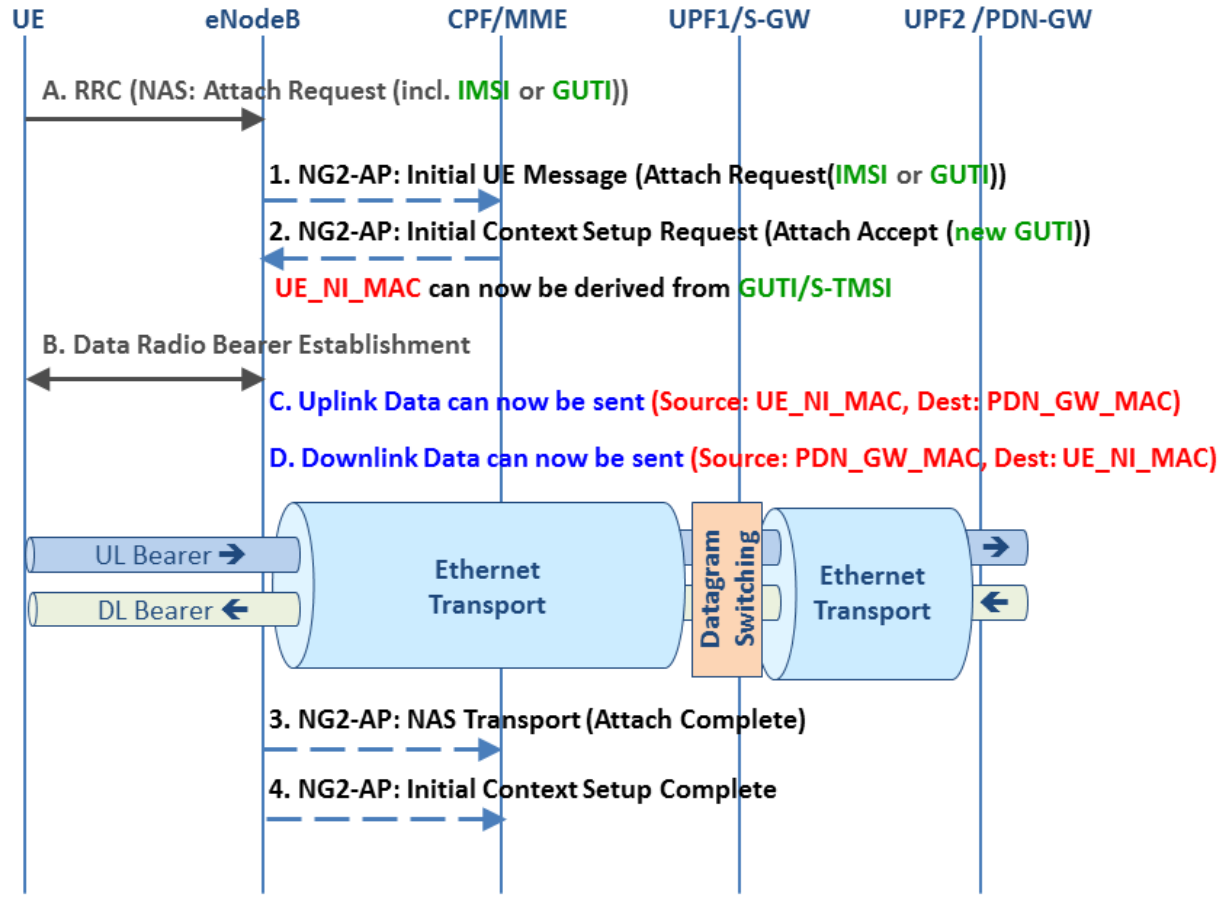


Figure 3-4: 5G Attach procedure using Ethernet based user plane

In total, 4 control plane messages are used between RAN and CN and inside the CN for the Attach procedure. For a service request procedure, no control plane messages would be needed for enabling the user plane data transport as the MAC address information is still available in the eNodeB and the UPFs.

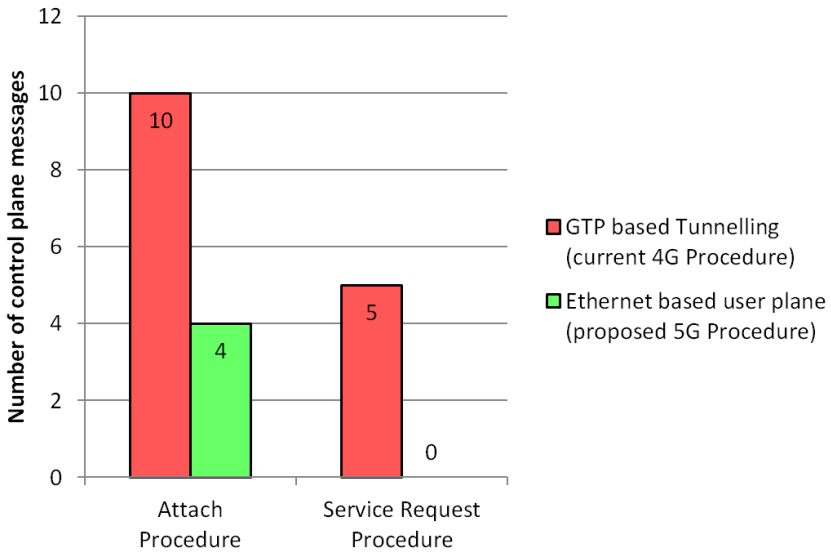


Figure 3-5: Required number of control plane messages

Figure 3-5 compares the number of required control plane messages with the GTP based Tunnelling approach as used currently in 4G. The figure shows that the connectionless packet switching approach using an Ethernet based user plane strongly reduces the number of required control plane messages compared to the connection oriented approach used in 4G based on the GTP tunnelling protocol.

3.3 QoS framework

In LTE, the CP functions in CN and RAN are responsible to differentiate the UP streams at the bearer level based on their QoS attributes. Compared to LTE, NR may support a more flexible/dynamic QoS differentiation with more criteria and finer granularity in the 5G RAN. In addition to this, NR shall support Quality of Service (QoS) and scheduling/priority mechanisms allowing consistent user experience over time for a given service everywhere this service is offered (as in LTE).

State of art

In LTE and the EPS, it is possible to establish different bearers with different QoS. The bearers can be of two different resources types, namely Minimum guaranteed bit rate (GBR) bearers Non-GBR bearers. The GBR bearers are typically used for applications such as VoIP while the Non GBR bearers for applications such as web browsing or FTP transfer.

The eNB ensures the necessary QoS for a bearer over the radio interface. Each bearer has an associated QoS Class Identifier (QCI), and an Allocation and Retention Priority (ARP). Each QCI is characterized by priority, packet delay budget and acceptable packet loss rate. The ARP of a bearer is used for call admission control i.e. whether the requested bearer should be established in case of radio congestion. The PDN Gateway is responsible for IP address allocation for the UE and QoS, see Figure 3-6 The PDN GW is responsible for the filtering of downlink user IP packets

into the different QoS-based bearers. This is performed based on Traffic Flow Templates (TFTs). The P-GW performs QoS enforcement for guaranteed bit rate (GBR) bearers.

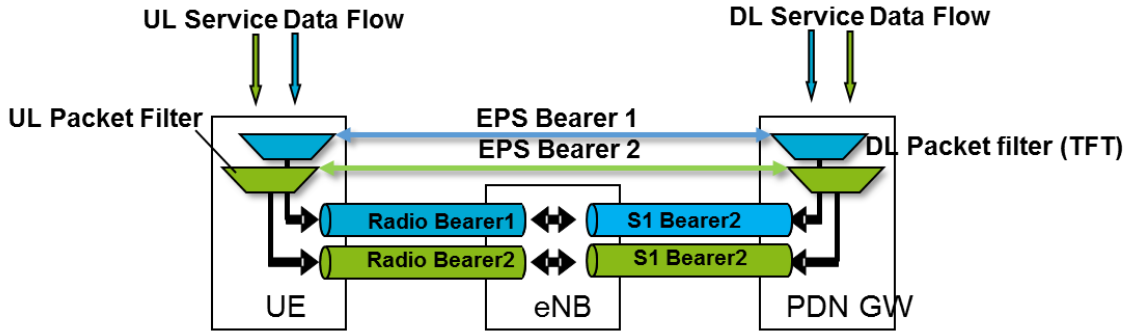


Figure 3-6 LTE QoS framework

The packets of an EPS bearer are transported by an S1 bearer between an S-GW and an eNB, and by a radio bearer between a UE and an eNB. The eNB stores a one-to-one mapping between a radio bearer ID and an S1 bearer to create the mapping between the two. IP packets mapped to the same EPS bearer has the same packet forwarding treatment, i.e. scheduling policy, queue management policy, rate shaping policy, RLC configuration. To provide different bearer-level QoS, a separate EPS bearer must therefore be established for each QoS flow. User IP packets must then be filtered (using TFTs) into the appropriate EPS bearers.

The TFTs use IP header information such as source and destination IP addresses and Transmission Control Protocol (TCP) port numbers to filter packets such as VoIP from web-browsing traffic, so that each can be sent down the respective bearers with appropriate QoS. An Uplink TFT (UL TFT) associated with each bearer in the UE filters IP packets to EPS bearers in the uplink direction. A Downlink TFT (DL TFT) in the PDN-GW is a similar set of downlink packet filters.

5G QoS framework

The basic idea is to make the NR QoS framework more flexible and dynamic with the option of finer granularity in the 5G RAN. One benefit with this is to support MC giving some flows over one AIV higher priority than the other. Another benefit is for services where packets are not equally important (due to coding for example), for example a video service. In that scenario, the eNB can assign the more important video packets with higher priority and the other video packets to default priority. For some cases, it is beneficial for the RAN to aggregate flows using same QoS into one single data radio bearer, e.g. due to efficiency reason. Note that dynamic QoS for traffic steering is analysed in [MII17-D52] and QoS from a UP aspect is covered in [MII17-D42].

One option to enable this is to avoid the one-to-one mapping of the data radio bearer (DRB) in the access network (eNB to UE) and the “S1 bearer”. 5G shall instead support a solution where the eNB decide whether to use one-to-one mapping or to use several different DRBs with different QoS. Figure 3-7: depicts three different possible NR QoS framework solutions proposed.

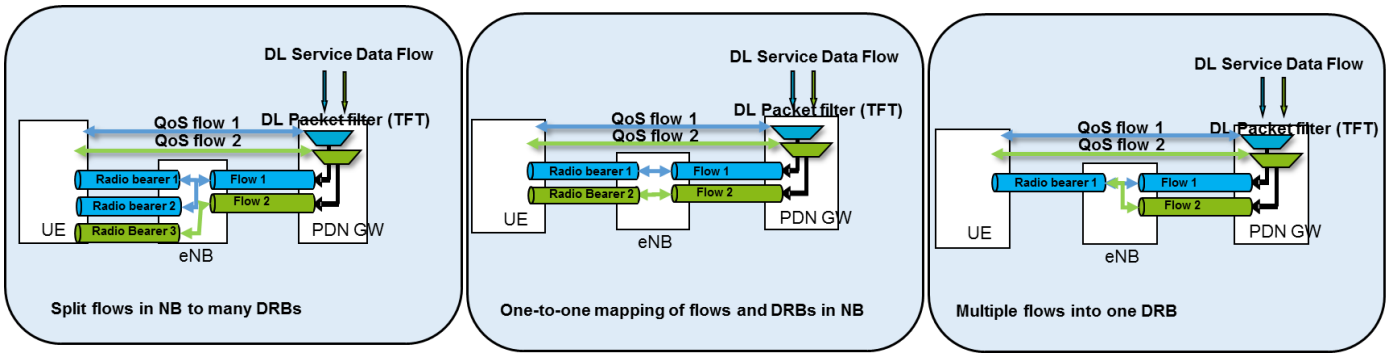


Figure 3-7: Possible NR QoS framework (only DL shown here)

The PDN gateway creates different QoS flows from the so called Service data flow based on the "5-tuple", i.e. the source and target IP address and port address (creating a unique "QoS flow"). The left-most solution shows the case when a "flow" (S1 bearer in LTE) is split into 2 DRBs. This solution is very flexible but rather complex since it requires the RAN to be able to distinguish and prioritize between the IP packets within a flow in some way. The middle figure shows the normal case, i.e. one-to-one mapping of flow and DRB. The right-most figure shows the case where multiple flows (using same QoS) are aggregated into one DRB. Note however that for all solutions the eNB must still support the overall QoS requirement from the PDN. The current situation in 3GPP is that one-to-one mapping and multiple flows (using same QoS) aggregated into one DRB are supported [3GPP17-38804].

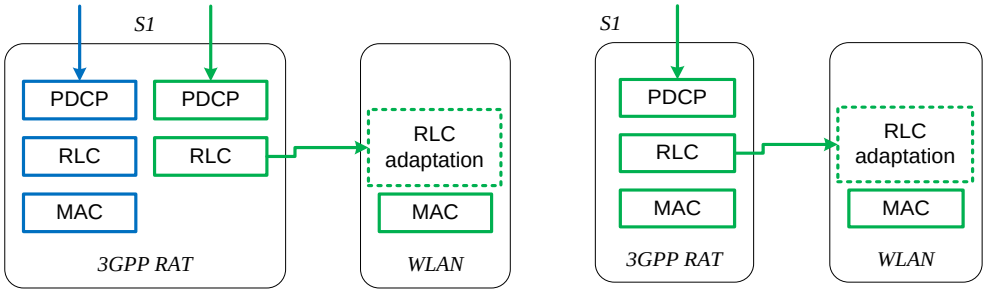
3.4 Integration with WiFi

The integration of 5G and WiFi systems can be useful for increasing the user throughput and to offload traffic from the cellular network. This section considers the integration of Wifi and 5G on spectrum below 6 GHz and compare solutions of tight integration (at lower layers) with loose integration (at higher layers). It is worth noting that the focus is on inter-operation between 5G and WiFi. There are also another options for integrating 3GPP and WiFi systems such as MPTCP (Multi-Path TCP) which is not considered in this section.

Solution 1: Tight 3GPP-WiFi integration at low layers

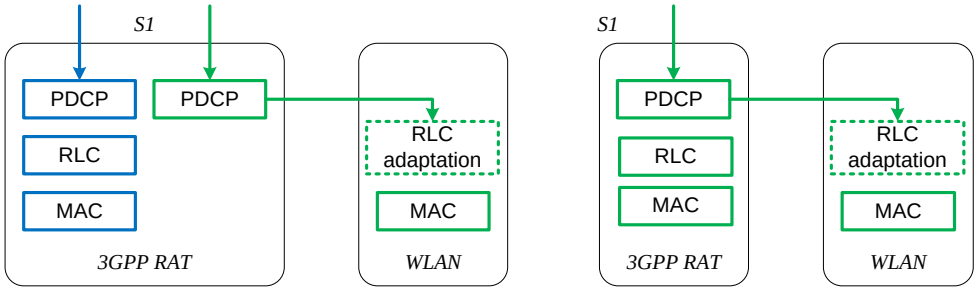
In this scenario, the packets are sent to WiFi or 3GPP cells on a rule that depends on the instant state of the two systems. The WiFi-AP needs thus to be directly connected with the base station or be collocated with it. The BS will be anchor for Control and User Plane and will be also responsible for traffic steering between two accesses. This implies that it has knowledge of the state of the two systems (e.g. numbers of users and their radio conditions). The integration itself could be realized on different protocol levels e.g. RLC or PDCP layer. To avoid IEEE 802.11 standard impact, the only reasonable approach is to analyze integration above the MAC layer of WiFi. RAN level integration should improve network utilization, user experience and provide more control for operators. Possible architectures for solution 1 are presented in Figure 3-8: and Figure 3-9:. There is no proposal for a MAC layer split because the 802.11 MAC is different from 3GPP

LTE MAC so such an architecture (resembling LTE Carrier Aggregation) is not possible between these different technologies.



a) b)

Figure 3-8: 3GPP-WiFi integration on RLC, a) no bearer split, b) with bearer split.



a) b)

Figure 3-9: 3GPP-WiFi integration on PDCP, a) no bearer split, b) with bearer split

Though the bearer split option might be more complicated because of the need to schedule packets from a single stream to multiple RATs it has the benefit to achieve higher user throughput. Because of the tight integration of both technologies in this scenario (together with very strict synchronization) bearer split is technically plausible and realizable in a real network implementation. So, in order to maximize the user experience the preferred architectures are options b) in these figures (similar to LTE Rel-12 Dual Connectivity feature).

Solution 2: Loose 3GPP-WiFi integration on a higher layer

In this scenario, the multi-homing (i.e. when a UE is using more than one access) policy is setup based on the average state of each system. Integration at the UE and RAN side would happen above PDCP protocol layer (i.e. PDCP SDU). The UE may receive user packets from the 3GPP radio side as well as from the WiFi and these are forwarded directly to upper layers. In this scenario, higher layers decide whether to switch or split the Data Radio Bearer to transmit over

3GPP radio and/or WiFi thus they need to have the knowledge of the average state of each system. Possible architectures for solution 2 are presented on Figure 3-10:

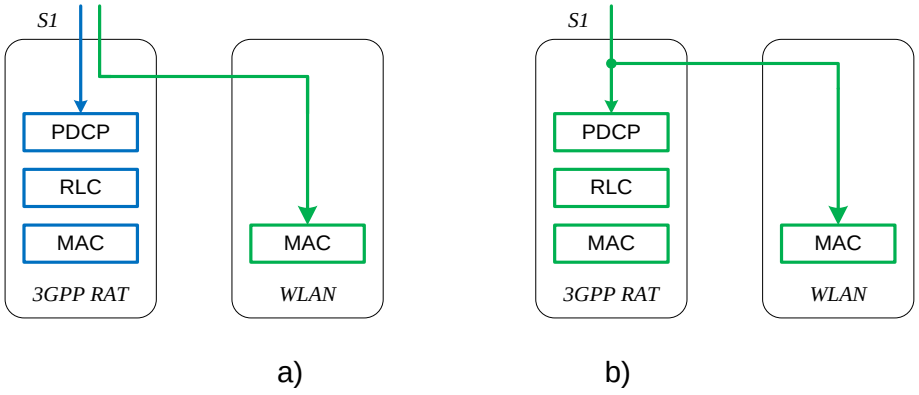


Figure 3-10: 3GPP-WiFi integration above PDCP, a) no bearer split, b) with bearer split.

In this scenario, there is no need for tight integration between 5G and WiFi nodes and they do not need to be collocated. This is similar to the current LTE Rel-13 LWIP solution (LTE/WiFi Radio Level Integration Using IPsec Tunnel). In this scenario there is an IP sec tunnel between a WiFi termination node and a UE connected to WiFi so the WiFi network does not even need to be aware of the ongoing aggregation with a 3GPP access. Because of no tight coordination between technologies the bearer split scenario is almost impossible to implement, therefore the architecture from Figure 3-10:-a) is the most plausible to realize in real network environments. We note here that this integration is possible also using different mechanisms such as Information Centric Networking (ICN) paradigm. In this case, the UE controls the multihoming approach by sending notification packets on the two air interfaces (5G and WiFi). Again, the bearer split is the best option here since the UE has knowledge of the long-term throughput on both systems and can split the data flow into two flows depending on this average state.

Simulation results:

In order to compare the performances of the two solutions, we consider a system where each 3GPP eNodeB coverage area contains 3 WiFi APs (IEEE 802.11n). We consider a dense network deployment with a cell radius of 350 meters and an operating frequency of 1800 MHz. We consider a homogeneous traffic over the coverage area, with 50% of UEs multihoming capable. A user that is not multi-homing capable connects to the system that offers the best signal quality.

We implemented the two above mentioned solutions. For solution 2 (3GPP-WiFi integration above PDCP) two flavors are tested. The first one is “peak rate maximization”, where the split of the bearer is based on the peak rate achieved on each of the air interfaces, while the second flavor is “network-assisted”, where the split considers also the load on each system to compute an estimated average user throughput. Figure 3-11: illustrates the average user throughput, for 3GPP only users, for WiFi only users and for multihomed (MH) users. We first observe that the network assisted strategy enables and 3GPP-only users to achieve higher throughput, especially at higher traffic loads. Yet, WiFi-only users have a slightly lower throughput with the network

assisted strategy than that obtained with the peak rate maximization. This difference of performance relates to the resource allocation strategy. While peak rate maximization uses throughput perceived by the users for each access type, network assisted strategy is more robust and takes into consideration the traffic intensity in each system. Consequently, we observe that multihomed users achieve higher throughput with the network assisted strategy when compared with single-homed users (3GPP-only and WiFi only). This is also the case for the peak rate maximization strategy but not at high traffic loads because this strategy does not take into consideration system's traffic intensity.

Figure 3-12: compares solution 2 (network assisted) with solution 1 (global). It can be seen that multihomed and 3GPP-only users achieve a better performance with global proportional fairness scheduling strategy whereas WiFi-only users performance becomes slightly lower than that achieved with network assisted strategy. The difference of performance relates to the precision of the allocation strategy: Solution 1 performs an instantaneous resource allocation, while solution 2 with network assistance uses average values for resource allocation decision. In contrast, the advantage of network assisted strategy, is the computational simplicity when compared with the complexity of the tight integration strategy.

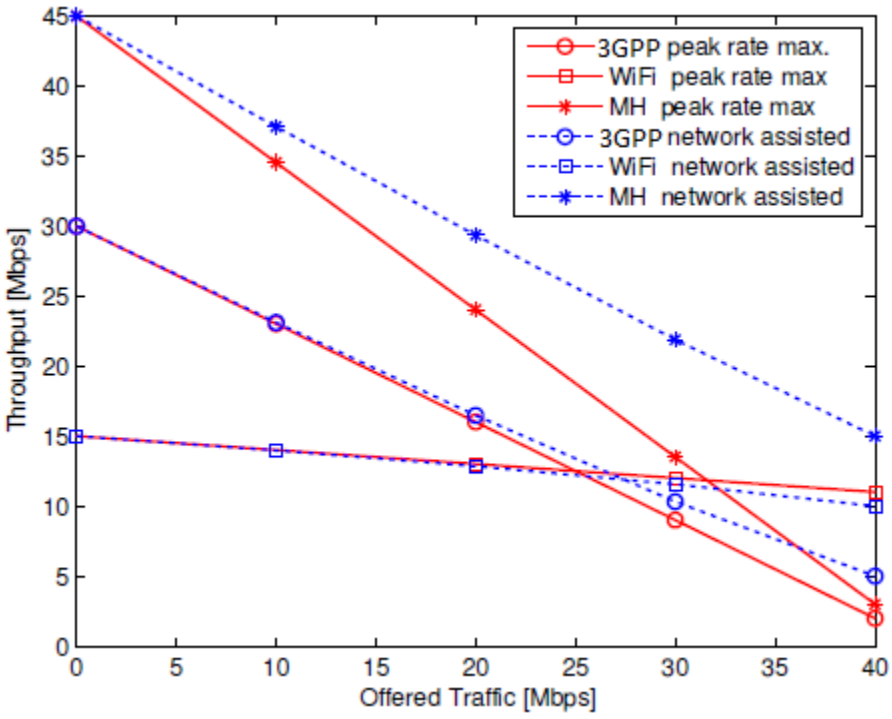


Figure 3-11: 3GPP-WiFi integration performance (solution 2).

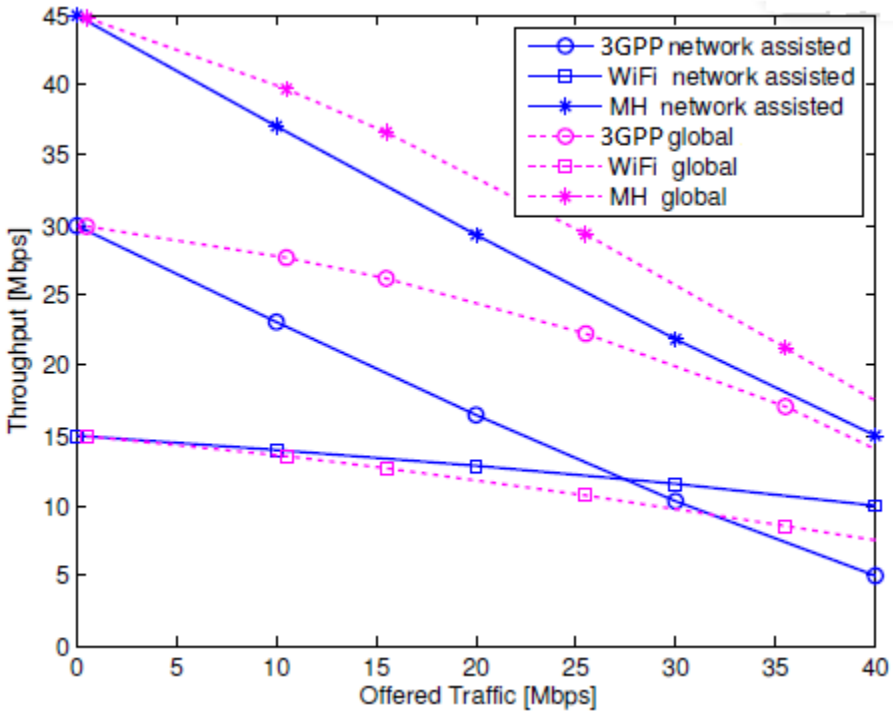


Figure 3-12: Comparing 3GPP-WiFi integration solutions.

Conclusions

As outlined in the above, both solutions have their pros and cons. The table below illustrates the pros and cons of the different solutions.

	Pros	**Cons**
Tight 3GPP-WiFi integration	1. Complexity of the WiFi network is hidden from the 3GPP system perspective 2. RAN manages resource utilization 3. Possibility to aggregate full throughput of 3GPP RAT and WiFi	1. Required ideal-backhaul between eNodeB and WiFi (low latency connection) 2. Impact on RAN specification 3. Required Adaptation Layer in WiFi AP (upgrades to WiFi deployment needed)
Loose 3GPP-WiFi integration	1. Limited impact on 3GPP specification 2. Possible deployment on non-ideal backhaul 3. Can work on legacy WiFi deployment	1. Not effective utilization of radio resources 2. No control of the traffic split in 3GPP RAN 3. Limited control of the WiFi AP (WiFi can be transparent)

		4. Bearer split almost impossible to implement (lower max throughputs than in scenario 1)

Solution 1 can achieve higher throughputs but is reliant on synchronization between both radio technologies and changes to current WiFi deployments. Scenario 2 on the other hand is a more over the top implementation of integrating 3GPP and WiFi RATs as it is done on higher layers. This means that the WiFi layer can be transparent and does not need to be aware of the ongoing integration with the other radio technology. But in this case, it is almost impossible to aggregate a single stream of data in both RATs which implies that the throughput could never reach the one achievable in scenario 1. Therefore, the scenario and architecture has to be chosen depending on the use case and required performance and implementation complexity levels.

3.5 Asynchronous spectrum management

From a mobile operator perspective, the current radio resource configuration in terms of spectrum allocation and usage (as well as the related radio technologies using such spectrum) can be considered, in principle, static. In other words, usually, a mobile network operator plans its network on the basis of some assumption/estimation on the traffic, dimensioning the network accordingly (e.g. number of sites, sites locations, bandwidths, etc.). Such network configuration is usually maintained as planned during all the network operating period, demanding to the RRM algorithms the management of the cell selection and re-selections (i.e. mobility), channel assignments, handovers, etc.. As depicted in Section 2.5, on top of RRM, SON functionalities can be introduced providing a network intelligence, automation and network management features in order to automate the configuration and optimization of the mobile radio access networks through inter-radio access technology operation, enhanced inter-cell interference coordination, coverage and capacity optimization and energy efficiency. Notwithstanding the flexibility introduced by the mentioned functionalities, taking into account the continuously increasing and dynamic (in time and space) traffic demand, the different type of services that can be provided in a network, as well as the multitudemultitude of different radio access technologies, the actual (semi)static network management approaches may result, in certain circumstances, in an inefficient radio resource (e.g. spectrum) utilization. On this basis, new management schemes for dynamic radio resources reconfiguration may be adopted to facilitate a more efficient use of the spectrum resources managed by an operator. In fact, exploiting such schemes, a network operator owning two or more radio systems could utilize the opportunity to dynamically and jointly manage the resources of the deployed radio systems, in order to adapt the radio resource configuration to the dynamic behavior of the traffic and to maximize the capacity where needed. Additional short-term spectrum management solutions are investigated in the METIS II deliverable D5.2 [MII17-D52].

In the context of the SON architecture presented in Section 2.5, an AIV reconfiguration management functionality can be introduced in the CPM framework as the entity in charge of the radio reconfiguration (e.g. spectrum) management with the goal of self-adapting towards an optimal mix of supported AIVs and frequency bands [MII16-D22]. This functionality may act on the basis of different input parameters, such as the available resources (radio and hardware), the traffic demand, the capabilities of the UEs within the NW (supported AIVs, frequency bands, etc.), and the requested services (e.g. bandwidth and QoS). In addition, this functionality could exploit a collaborative AIV reconfiguration management scheme, where the decision making functions are shared among different network nodes. As a high level example, Figure 3-13 depicts a generic scenario that includes intra-system and inter-system reconfigurations and that implicates modifications to both radio and hardware resources for the involved AIVs.

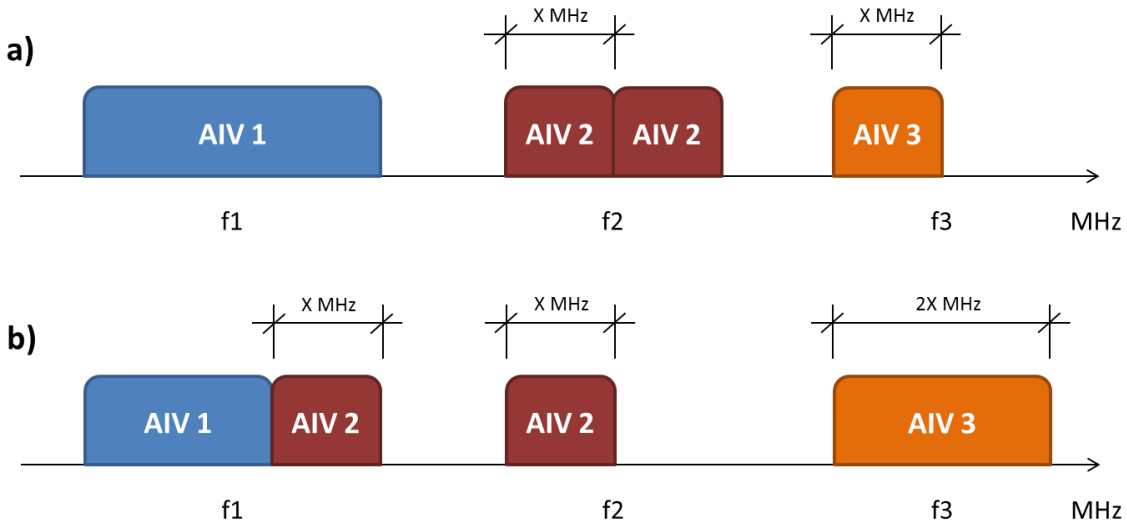


Figure 3-13: Examples of AIV reconfigurations

In relation to the current 3GPP activities on 5G, the generic concept depicted above, can be applied in the context of co-existence and interworking between the NR (New Radio) and legacy RATs currently part of a study item work [3GPP17-38913]. In an intra-operator domain, it is under discussion that the NR should be able to support flexible allocation of resources (e.g. time, frequency) between the NR and the legacy RATs (e.g. LTE) operating in the same block of spectrum (with possible bandwidths overlap). Resource allocation granularity in the time/frequency domain, as well as the potential guards between NR and LTE resources are still to be determined by the study on the NR. The NR should be able to use these resources at least for downlink, uplink and eventually sidelink and it should work whether legacy RATs are supported by the same base station as the NR or the two RATs are supported by different base stations. On such basis, such a flexible allocation of resources may also enable a smooth introduction of the NR in the same band used by a legacy RAT; for example, the band allocated to a legacy RAT can be progressively reduced (by steps of 5 MHz) in order to make band available for the

allocation to the NR. Focusing on LTE as a legacy RAT, the coexistence of NR and LTE can be categorized into two main categories [3GPP17-R11700031]:

- FDM (Frequency-Division Multiplexing): in this case, NR and LTE have no bandwidth overlap and to fulfill adjacent channel coexistence requirements, guard band between NR and LTE are needed. Additionally, bandwidth adaptation mechanisms can be used to balance the traffic loads of NR and LTE as well as by adopting cell (de)activation mechanism.
- TDM (Time-Division Multiplexing): NR and LTE have bandwidth overlap. Because of such tight co-channel coexistence, special mechanisms for interference management based on dynamic sub-frame allocation are needed for both NR and LTE.

In addition to the NR-LTE coexistence in both DL and UL carriers, NR and LTE can only share the same UL carriers but have separate DL carriers. The approaches under discussion described above can be further categorized in static, semi-static and dynamic [3GPP17-R11700841]:

- *StaticStatic FDM*: spectrum partitioning between NR and LTE (e.g. 5/15, 10/10, or 15/5 MHz partitioning assuming 20 MHz bandwidth) can be adjusted based on UE penetration.
- *Semi-static FDM with CA (Carrier Aggregation)*: LTE has static BW (BandWidth) allocation as anchor for LTE and NR (e.g. 5 MHz) and the remaining BW is allocated to LTE SCell and/or NR as needed. Another possibility would be that LTE and NR has static PCell (Primary Cell) BW allocation (e.g. 5 MHz). Remaining BW allocated to LTE SCells and/or NR SCells as needed.
- *Semi-static TDM*: utilize LTE DL MBSFN subframes and resources with unused UL subframes to schedule NR.
- *Dynamic Resource Sharing*: NR utilizes unused LTE resources dynamically at PRB level in frequency and subframe level in time.

In such a context, the Network Management and Orchestration framework presented in [MII16-D22], may find an application in the NR/LTE co-existence management in which the reconfiguration of the radio resources (e.g. change of channel bandwidth, activation of a novel AIV in a different frequency band, etc.) of NR and/or LTE could be performed by the AIV reconfiguration management functionality on a slower time scale (e.g. on the order of hours), while the allocation of the specific channel resources (i.e. PRBs) to NR and/or LTE could be handled in the context of the agile RM (Resource management) framework presented in Section 6.2 of [MII16-D22]. The presence of an Orchestrator entity managing the interaction between such two levels of resource management would add more efficiency as well as flexibility in the overall network control and management.

4 State Handling

4.1 Introduction

In 5G, the system should give the end user the perception of being always connected, thus the access to the network and state transition to connected state should be instantaneous from the end user perspective [NGMN15], in the order of 10ms. Hence, to tackle this challenge, the 5G systems must adopt connectivity with flexibility in its configuration to system access [NOK14].

The novel RRC state model was introduced in [MII16-D61] and further adopted to discussions in the 3GPP standardization working groups after the 3GPP RAN#71 meeting approved the New Radio study item [3GPP16-RP160671]. This chapter describes the recent advances in the proposed RRC state model compared to the 3GPP status [3GPP17-38804]. The chapter starts with some background and the lessons learnt from previous systems to motivate the introduction of a new RRC state model. In this context, the always-connected low activity UE refers to situation where the UE starts an inactivity period and it remains connected to network and keeps parts of the RRC context to later resume the connection with minimum signaling latency and overhead. The new RRC state is herein called *RRC Connected Inactive* [SMS+16] or “RRC_INACTIVE” in 3GPP.

The take-away of this section is to understand the proposed state model and cover the identified characteristics of the new proposed state called RRC Connected Inactive. The proposed state model [MII16-D61] and the agreements in 3GPP [3GPP17-38804] are well in line and optimize the power consumption of mobile devices during the low activity periods while minimizing the latency for the first packet transmission from the UEs to the network. The new state model is shown to be configurable based on different aspects of use cases, device capability, access latency and security requirements, privacy, etc. The section derives mobility-related signaling diagrams for RRC Connected Inactive. Finally, the RRC Connected Inactive state is shown to support small data transmission without state transition to Connected state. This is especially useful for the frequent small data traffic profiles typical for mMTC and IoT devices. Finally, the RRC Idle state in 5G is used for example during UE power-on, inter-RAT cell selections and fault recovery.

4.2 Background on State Handling

4.2.1 Introduction

An overview of existing RRC states and state transitions specified by 3GPP is shown in [3GPP16-36331], which also illustrates the mobility support between E-UTRAN, UTRAN and GERAN.

4.2.2 RRC States in HSPA

The RRC states in HSPA depend on the physical channels which are allocated to the UE and the transport channels that can be used. Mobility procedures and the UE activity define which RRC

state is in use and how the state transitions are used by UE or configured for UE. In HSPA the number of RRC states is quite high (w.r.t LTE) due to dedicated characteristic features per state. For example, CELL_PCH and URA_PCH have almost identical characteristics with UE monitoring the PCH (Paging channel), except that the location of a UE is tracked at the cell level in the former case, whilst tracked on the URA (UTRAN Registration Area) level in the latter.

Table 4-1: RRC states in HSPA

HSPA State	Mobility Procedure	Monitoring Dedicated Physical Channels	DL Channel Monitoring	Location Update	Uplink Activity Allowed	Storage of RAN Context Information
CELL_DCH	Network controlled handover	Yes	Continuous (DCH)	Active set update	Yes	Yes
CELL_FACH	Cell selection & reselection	No	Continuous (FACH)	Cell update	Yes	Yes
CELL_PCH	Cell selection & reselection	No	Discontinuous with DRX (PCH)	Cell update	No	Yes
URA_PCH	Cell selection & reselection	No	Discontinuous with DRX (PCH)	URA update	No	Yes
IDLE	Cell selection & reselection	No	-	-	-	No

4.2.3 RRC States in LTE

The two RRC states in LTE, RRC_IDLE and RRC_CONNECTED, reduce the complexity of mobility and connectivity procedures arising from having multiple states as in HSPA systems. The RRC_IDLE minimizes the UE power consumption, network resource usage and memory consumption while the RRC_CONNECTED was introduced for high UE activity and network controlled mobility. The state transition from CONNECTED to IDLE and vice-versa requires considerable amount of signaling to setup the UE's AS context in RAN and introduces delay beyond the 5G CP latency requirement of 10ms. Target 5G CP latency value of 10ms can refer for example to 3GPP Rel-13 LTE Advanced 3GPP-36912] where the dormant-to-active state transition for synchronized UEs could achieve CP latency in the order of 10ms.

Table 4-2: RRC states in LTE

LTE State	Mobility procedure	Monitoring Dedicated Physical Channels	Allowed Mode for DL Channel Monitoring	UE Location Known on	Uplink Activity Allowed	Storage of RAN Context Information
RRC_IDLE	Cell selection & reselection	No	Discontinuous with DRX	Tracking Area list level	No	No
RRC_CONNECTED	Network controlled handover	Yes	Both continuous and discontinuous with DRX	Cell level	Yes	Yes

Smartphone applications and MTC devices have resulted the reconsideration of state handling efficiency in LTE systems. It is costly to keep all the UEs in RRC_CONNECTED state, as high user activity increases power consumption and mobility related signaling. On the other hand, the state transition from RRC_IDLE to RRC_CONNECTED, in the order of 50ms, cannot meet the 10ms latency requirements of 5G.

4.2.4 Control Plane Latency

The CP latency is defined as the time required for the UE to do state transition from the lowest power saving state to the state where the UE is ready for data transmission. In Idle state of LTE, the UE does not have an RRC connection, thus once the RRC is setup procedure is done, the UE transitions to connected state and is ready for data transmission. Transition time from Idle to Connected state takes around 50ms [3GPP15-36912]. A quick system access and fast transmission of the first packet could be achieved by maintaining all UEs in RRC Connected with discontinuous reception DRX optimized for RRC Connected state. However, the RRC Connected procedures are optimized for data transmission rather than low battery consumption so this becomes inefficient for slightly longer inactivity periods [LNM+13].

CP latency evaluation is done for 5G systems assuming that the UE context is stored in RAN node during the low activity state which includes the storage of AS and NAS security context information, at least some parts of data and signaling radio bearers and the low activity state mobility and the UE reachability functions are configured. With such assumptions, the UE connection can be inactivated and activated, which will reduce the CP latency. It can be also assumed that 5G sub-frames will allow further division to shorter UE specific slots in time on the OFDM symbol level and the UE and network nodes will have better processing capabilities. Thereafter it can be expected that there is significant reduction in the overhead time required to change from the low activity state to fully connected RRC state in NR systems.

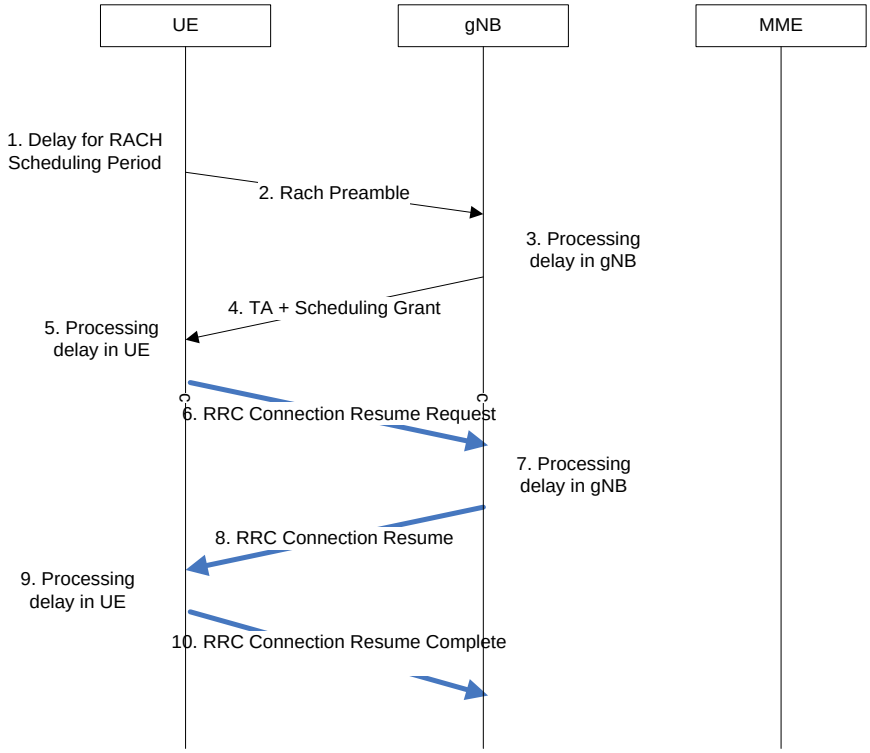


Figure 4-1: Signal flow diagram in the new state for NR (without context fetch)

The CP latency can be estimated with following assumptions:

- DL scanning and synchronization + broadcast channel (BCH) acquisition.
- Random access (RA) procedure:
    - Average delay due to Random Access Channel (RACH) scheduling period.
    - RACH Preamble transmission
    - Preamble detection and transmission of Random Access response
    - UE processing delay (decoding of scheduling grant, timing alignment and identifier assignment + encoding of RRC CONNECTION RESUME REQUEST)
- RRC connection establishment/resume
    - Transmission of RRC CONNECTION RESUME REQUEST
    - Processing delay in BS (L2 and RRC)
    - Transmission of RRC CONNECTION RESUME to UE with e.g. UE ID
    - Processing delay in the UE (L2 and RRC)
    - Transmission of RRC CONNECTION RESUME acknowledgement to BS

The transmission delay over the air interface can use both normal TTI sub-frame of 1ms and as an example shortened TTI length. Further for lower latency it can be assumed that the 5th generation UEs and network nodes will have better processing capabilities.

Based on the calculations in [MII17-D23, Table 3.1] for transition from RRC Connected Inactive to RRC Connected, CP latency as short as 7.125ms can be expected for 5G, when the sub-frame of 0.25ms was selected. It should be noted that further reduction of CP latency for 5G can be obtained if e.g., sub-frame duration of 0.125ms is used. This is on condition that the UE stays in the same serving cell or performs cell updates after each cell reselection, thus not counting the extra delay due to the UE context fetch procedure from the last serving network node. The UE context fetch procedure can obviously be avoided, if the UE is continued to be served by one of the cells of the network node during the low activity period.

4.3 RRC Connected State

4.3.1 Introduction

Mobility refers to the system's ability to provide seamless service experience to users that are moving. For active users, the mobility procedures offer services such as voice or real time video connections where the connections can be maintained active for all mobility profiles, even when the user is moving at very high speeds.

In 5G the number of use cases is significantly broader compared to LTE and this has raised new requirements also for mobility. The 5G use cases show that 5G networks need to support a growing set of static and nomadic users and devices. On the other hand, mobility functions inside the 5G RAT can be simplified for UEs operating on a limited geographical area and this may help decreasing the cost of infrastructure and devices. 5G mobility solutions could limit the mobility support for some devices and services and provide simplified mobility on demand to those devices and services that need it only on the local scale.

4.3.2 Characteristics of Connected State

Main characteristic of connected state is the RRC connection between the UE and the Network and allocation of logical dedicated unicast resources for the transfer of control plane signaling or user plane data in uplink or downlink.

The UE has AS context in RAN and RAN knows the cell where the UE is located. Thereafter RAN can transfer unicast data to/from UE without RRC state transition related extra signaling.

The network controls the mobility by performing handovers and cell changes and the UE location is known at the cell level. UE will be listening to control channels and perform measurements and measurement reporting according to configuration from the network to assist network mobility functions and procedures.

The UE monitors the paging and notification channel from RAN and will monitor system information broadcasted by the network. System information may be complemented with

dedicated UE information and parameters. As part of the active data connection, the UE provides connection feedback information to the network such as channel quality and channel state information.

Apart from DRX periods, the UE has to be awake at all times to decode the incoming downlink control information and user data, as the data in the downlink may arrive at any time based on scheduling from network.

4.4 RRC Connected Inactive State

Connectivity solution where the UE is kept 'always ON' from the Core Network (CN) perspective, is considered to achieve a seamless UE state transition between low activity state and high performance state in 5G systems. Once the UE is registered to the 5G network, the connection to the CN is kept alive. However, the RAN can suspend [3GPP15-22720] the RRC connection during inactivity times. The RAN has also the opportunity to configure differently the behavior of UEs with different service requirements during inactivity times.

4.4.1 Radio Resource Control State Transitions

Although the UE is in ECM connected state from the CN perspective, the RAN has the opportunity to suspend the RRC state of the UE during inactivity periods. From a RAN perspective, the UE can be in RRC connected, RRC Connected Inactive or RRC idle state. When inactivity is detected, the UE may request that the network suspends the RRC connection. This request may be based on a configured timer. Alternatively, the RAN may suspend the RRC connection after the data buffers are empty or if there is a temporary inactivity detected. The RRC idle state may be rarely used, for example, as a recovery state when RRC resume fails.

The UE mobility is controlled by the network during RRC connected state. However, during RRC Connected Inactive state, it is envisioned that the mobility of the UE is controlled by the UE with the assistance of the network. In RRC Idle state, i.e. in a kind of recovery state, the UE mobility using cell reselections is autonomously controlled by UE. Lightweight procedures called RRC suspend and RRC resume are respectively used to resume and suspend RRC connection according to Figure 4-2. The RRC suspend message may contain service tailored configuration in order to address UEs with diverse service requirements [SMS+16].

When the network commands the UE to Connected Inactive state, the last serving 5G-NB sends an RRC Connection Suspend message to the UE. The message that contains (at least) Resume Identification (ID, in this case the Last 5G-NB ID), Connected Inactive state related timing Information (e.g. Registration period), up-to-date TA List in which UE is allowed to move without TA update and Security Information for UE identification while re-connecting to the network.

Active connection in RRC Connected state is needed again when an application needs to send data. The UE is already connected so it will reconnect to the network via its current 5G-NB cell and send the RRC Connection Resume Request message to the 5G-NB including (at least) UE ID, Resume ID, Connected Inactive state related timing Information (e.g. time spent in inactive

state), and Security Information to verify the UE context. The 5G-NB responds to the UE with the RRC Connection Resume Complete message and UE is back to CONNECTED state.

A connection failure during the RRC Connected Inactive period can happen for example due to failed cell reselection or if the cell update to RAN was not acknowledged back to UE. Also, RAN can detect and assume a connection failure or UE may have been powered off if RAN has not received any location update information from UE within the maximum reporting time period.

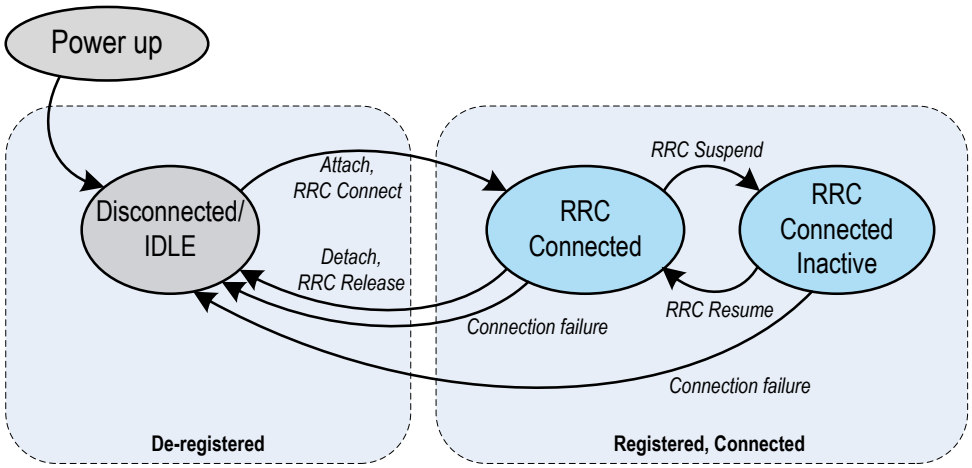


Figure 4-2: 5G UE RRC state transitions

4.4.2 Configurability of RRC Connected Inactive State

The need for configurability of the RRC Connected Inactive state is motivated by 5G use cases which have highly diverse, and sometimes contradictory requirements in terms of reliability, mobility, latency, bandwidth, security and privacy, battery life etc. For example, the E2E latency requirement varies from <1ms for use cases with ultra-low latency requirement such as autonomous driving, to latencies from seconds to hours in use cases under the category of 'Massive low-cost/long-range/low-power MTC' applications. The battery life requirement is irrelevant in some use cases like autonomous driving, where the device can get unlimited energy from the car, whereas the battery life requirement for battery operated devices ranges from three days for smartphones to up to 15 years for a low-cost MTC device.

Allowing a device to use a specific RRC Connected Inactive state configuration enables flexibility to the state handling mechanism. The solutions in [3GPP16-22891] for state transition optimization for ultra-low complexity, power constrained, and low data-rate Internet of Things devices can be taken as an example of network slice specific state handling configuration. However, the solutions described in [3GPP16-22891] may not necessarily be applicable for all use cases, e.g. autonomous driving of vehicle.

Assuming that part of the RAN context is available at the network and the UE, some of the potential RRC Connected Inactive state configuration options include:

- Mobility/location tracking management configuration: RAN based mobility and location tracking, single/multiple cell-level tracking
- Measurement configuration: Measurement configuration for cell reselection, camping, etc. taking into account the existence of multiple AIVs
- Camping configuration: Single/multiple-RAT camping, capacity based camping, etc.
- State transition/system access configuration: State transition and RACH access optimizations
- Synchronization configuration: DL and/or UL synchronization.

The RRC connected inactive state of a UE can be configured based on the characteristics of the service(s) provided to the UE if such information is available at the network. According to Figure 4-3, a service could be characterized based on its requirements, which are indicated to the network as part of the Suspend Request. Such characteristics could be for example mobility, security & privacy, reliability, bandwidth, latency, battery life, etc. The configuration of the RRC Connected Inactive state is included in the Suspend message. If the UE has multiple services or purposes, e.g. a device with multiple concurrent services, then the configuration might be done based on the service with the most stringent requirement.

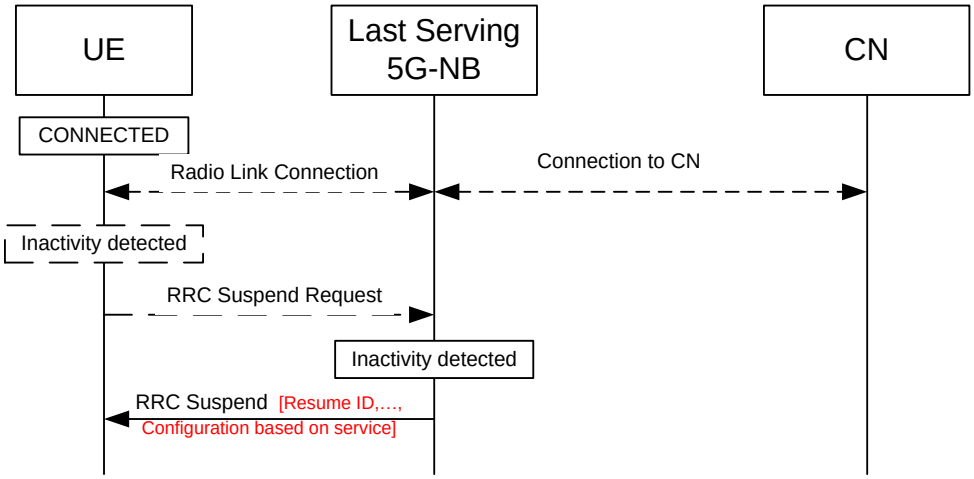


Figure 4-3: Configuration of a RRC Connected Inactive state

4.4.3 RRC Connected Inactive for Small Uplink Data Transmission

The number of system access events and small data traffic increases significantly with the growth of the always-on applications, which need to be constantly reachable by the network. Keep-alive messages generate autonomous data traffic of UP data packets between the UE and the network to maintain IP connection without user interaction with device. The MTC devices, on the other hand, generate infrequent small burst transmissions of the order of one or two messages per day. This kind of traffic is generally low in volume and consists of packets arriving at the network in a bursty way, and can be widely dispersed in time [3GPP11-36822]. The MTC and IoT type of

devices and applications would benefit from a minimized signaling overhead to enable efficient short message transmission. Minimizing the UE state transitions the can also reduce the signaling exchange with network for attach, link establishment and scheduling before the uplink transmission can take place.

In [3GPP16-R1163961] some potential physical layer techniques have been discussed, where autonomous, grant-free and/or contention based non-orthogonal multiple access will be studied for UL transmission for the MTC use case. A contention based UL transmission scheme has the potential to reduce signaling overhead, latency and power consumption. The proposed RRC state model can fulfil the requirements of various always-on use cases. In these applications, a low activity device in RRC Connected Inactive state could be configured for contention based UL transmission without the need to establish the connection and transition to RRC_CONNECTED state. Therefore, it could be useful to keep low activity UEs in RRC Connected Inactive and use the configurable properties of the state based on the service requirements or device type for small data UL transmission. This approach is captured in 3GPP report [3GPP17-38804].

4.4.4 Benefits of Connected Inactive State

The number of RRC states in LTE is two, which reduces the complexity compared to multiple states in HSPA addressing different use cases and low activity periods with UE location tracking. In LTE RRC_IDLE was optimized to minimize the UE power consumption, network resource usage and memory consumption, while the RRC_CONNECTED was developed for high UE activity and continuous data transmission with DRX used as a power saving function when the UE is connected to network. The RRC_IDLE to RRC_CONNECTED state transition requires considerable amount of signaling to setup the UE's access stratum context. That is, the state handling in LTE works for connectivity where frequent state transitions are not required.

The proposed state model design for 5G learns from the pros and cons of state handling design in the existing technologies and take into account 5G use cases and their corresponding requirements. Some of the benefits of the proposed state model are:

- can keep the UEs always connected from CN perspective.
- enables minimum number of RRC states to avoid added complexity in the state model.
- provides fast and lightweight state transition between active data transmission and power saving, which can reduce CP signaling overhead for frequent state transitions.
- can fulfil the latency requirement of state transition for control plane.
- can support highly configurable procedures for contradictory requirements of various 5G use cases.
- allows network slice specific state configuration.

4.5 RRC Idle State

The need for Idle mode has been questioned in 5G as the "Always on Applications" used in the smart phones need to send and receive small packets frequently to keep IP connectivity open. Essentially the frequent connection requests are related to push services where the UE is

checking if there is any new information in their application server. Typically, this is UE initiated and happens in un-controlled way from the radio network perspective. Another problem from network perspective is the “heart beat” or “keep alive” messaging that may occur once per minute, or once every few minutes, and the amount of data is very small (<<1 Kbyte). These messages are used to keep the device connectivity towards network in RRC Connected state. METIS-II assumes that the UEs (including smart phones, MTC devices or other kinds of terminals) can be always connected to network and that the operational state transition in UE between inactive/connected and active/connected can provide power savings during inactivity and optimized performance during active state, and that the state transition is introducing minimum system access latency.

The new RRC Connected Inactive state has many features of the existing LTE IDLE state, such as low activity towards network and UE based mobility using the cell reselection procedure. Despite the enhanced features/functionality, the IDLE state in 5G mobile systems may be needed. One major reason is that the needed fault recovery mechanism will add complexity to the new Connected Inactive state. It is an essential requirement for the UE to be able to revert or fallback to a recovery state in case of sudden connectivity fault or network failure. The IDLE state in 5G can for example support the bootstrap procedures, initial PLMN selection, UE controlled mobility and core network based location tracking.

4.6 Inter-RAT state transitions with RRC Connected Inactive

It has been acknowledged that the evolution of LTE should be integrated to the 5G in order to possibly benefit from the widely-deployed coverage of E-UTRA LTE in the 2020 timeframe. Tight integration of LTE and new 5G AIVs should not introduce additional core network signaling complexity to the RRC state handling. A moving UE RRC connection may be suspended or inactivated and UE is using the RRC Connected Inactive state during the low activity period. With tight integration, the connection resumption or activation back to RRC Connected could be done based on the system which is able to provide better coverage or capacity according to Figure 4-4. That is, a connection inactivation in 5G may be followed by resumption in LTE when 5G coverage is not available in the same geographical area.

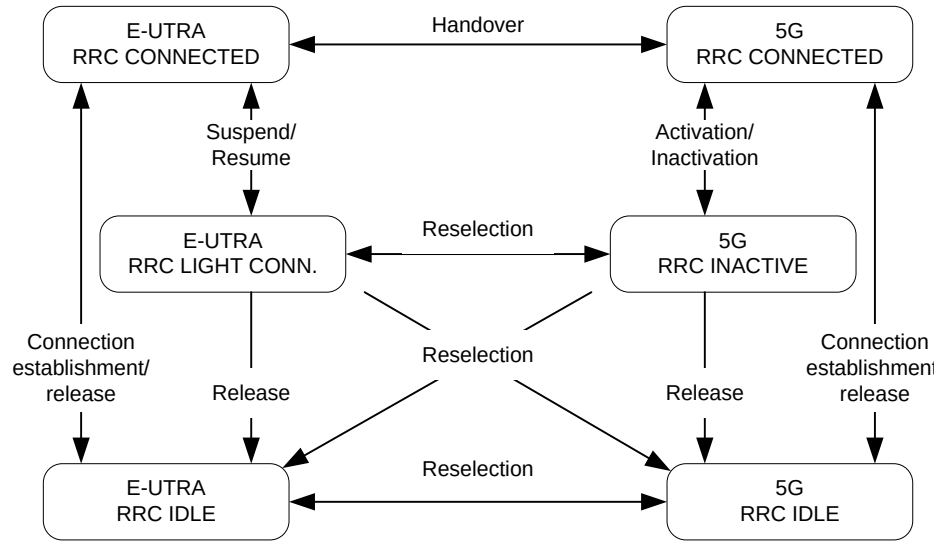


Figure 4-4: UE states and state transitions for NR and Inter-working with E-UTRA

4.7 Summary

This chapter has described the RRC state model for 5G and cover the identified characteristics of the new proposed state called RRC Connected Inactive. The new state model and proposed new state together optimize the power consumption of mobile devices during the low activity periods while minimizing the latency for the first packet transmission from the UEs to the network. The identified characteristics of the RRC state model are shown in the Table 4-3. The mobility and system access procedures of the new state model are configurable based on different aspects of use cases, device capability, access latency, power saving, security requirements and privacy.

Table 4-3: RRC states in 5G

5G State	Mobility procedure	Monitoring Dedicated Physical Channels	Allowed Mode for DL Channel Monitoring	UE Location Known on	Uplink Activity Allowed	Storage of RAN Context Information
RRC_IDLE	Cell selection & reselection	No	Discontinuous with DRX	Tracking Area list level	No	No
RRC_CONNECTED_INACTIVE	Cell selection & reselection	Configurable, yes/no	Discontinuous with DRX	RAN Tracking Area level	Configurable, Contention based UL data	Yes
RRC_CONNECTED	Network controlled handover	Yes	Both continuous and discontinuous with DRX	Cell level	Yes	Yes

5 Initial Access

Initial access refers to a set of CP functions across multiple layers of the RAN protocol stack (e.g. PHY, MAC and RRC) and, at some extent, the CN / RAN interface as in the case of paging and state handling. In LTE, some of these functions are synchronization (time and frequency, UL/DL), Cell Search, System information distribution and acquisition, Random access and Paging [3GPP17-36300].

This section relates to solutions tackling the random access, the paging, and the system information distribution using lean design. For the first aspect, the developed solutions aim at handling both the initial access bottlenecks due to massive connectivity and the service prioritization. The second problem that it is treated in this chapter relates to the signaling cost due to paging individual UEs, a process which is very costly in case of high moving devices. In the proposed scheme, the merits of the connected inactive states are being used for reducing the signaling cost in the RAN. Finally, for the system information distribution, the idea is to have a leaner design which will decouple the new system from the physical Cell ID, aiming at a higher level of future-proofness. The intention of such an approach is to avoid the need to transmit system information on a per cell basis, which is proven very inefficient and use self-contained transmissions where system information decoding is not scrambled with the cell ID.

The key takeaways of this chapter could be summarized in the following bullet-list:

- RACH needs significant modifications to enable massive access or service differentiation. This can be done using device grouping and preamble coding.
- RAN paging can offer significant signaling reduction in the context transfer and the paging channels load
- Lean design decouples the system from the physical Cell ID and reduces the need for constant transmission of signals – thus enabling reduction on the energy costs. With the presence of beamforming such schemes offer even higher energy gains.

5.1 Random Access Channel Solutions

In the following years the number of the MTC devices will increase significantly according to recent analyses [CISCO+16] . Thus, the networks and the corresponding signaling mechanisms that are designed to support mainly human-to-human communications (e.g., Telephony, SMS, Streaming services, etc) will not be able to handle this increase, thus requiring for innovative solutions [3GPP17-36300]. To efficiently support MTC, it is required to design new schemes that will lead to the reduction of signaling messages both in downlink and in uplink communication and avoid potential communication bottlenecks for a 5G operator in channels such as the random access and the paging.

In LTE, for accessing the network, the user equipment (UEs) follow the contention-based random access procedure, which occurs in every Random Access Opportunity (RAO) (Figure 5-2). However, such network designs are unlikely to be able to handle the MTC applications, where a large number of machines will attempt to transmit simultaneously small amounts of data. According to [ZN13] the collision probability for the RACH procedure will be almost certain (99.97%) for a cell with 1000 users and 30 ms packet arrival interval (with the 64 available RACH preambles).

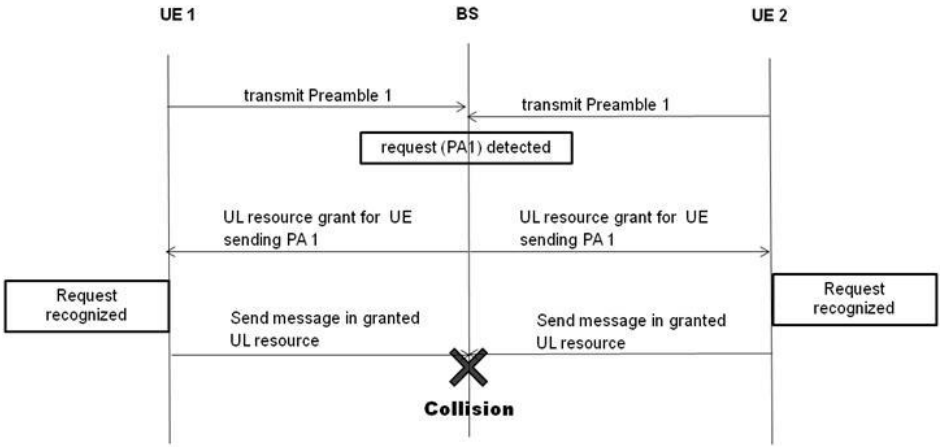


Figure 5-1: Collision in LTE random access

Up to now several schemes have been proposed in the literature for handling the RACH procedure. These schemes may be classified into two large groups, namely pull-based and push based [CKS+15]. Then, based on their special characteristics they may be further classified to other sub-categories, as shown below:

- Pull-based schemes
  - Access Class Barring where devices are categorized in classes and are specific time slots are statically allocated to each class
  - Back-off timer, where devices select to reattempt data transmission after an arbitrary time period
  - Dynamic RACH allocation, where the RACH resources used per time frame are adapted based on the identified collisions
  - RACH resource separation split, where the 64 preambles used for RACH are split to two groups; one solely used by UEs and one used either by typical UEs or MTC devices.
- Push-based schemes
  - Paging Control schemes where the push based procedure of paging is used to achieve Small Data Transmission (SDT) from the MTC devices

These schemes however, are designed mainly for prioritizing access based on the transmission requirements and are not, on the one hand, targeting the solution of the collision rate problem,

and on the other hand, are not focusing on 5G use cases (such as V2X, smart grid, mMTC) but rather focus on traditional use cases. Even in the cases where the solutions are applied for MTC scenarios, the number of the considered devices is rather small, thus making their applicability in scenarios where big number of devices is considered (e.g., [MII16-D11], [NGMN15]) highly questionable. Additionally, the sparse random access channel resources increase the latency to unacceptable levels for ultra-reliability scenarios.

In this section, we propose 2 new approaches to increase the RACH efficiency. The first concept is called Group based RACH in Section 5.1.1, tackles the massive connectivity problem using groups formed considering (apart from other metrics such as mobility, channel quality) the service requirements of the UEs whereas the second one, presented in 5.1.2, is called RACH Multiplexing aims at supporting the diverse access requirements of the users. A third subsection presents a summary of a mechanism described in D6.1 [MII16D61] related to Synchronization Sequences and Random Access in Higher Frequencies for the shake of completion.

5.1.1 Group based RACH

For the random access of vast amount of devices, a solution based on the grouping of the devices seems to be appropriate, since instead of having all the group members to proceed in random access using one of the 64 preambles when they have to transmit we could aggregate the transmission requests and only one device (the group head) will perform the RACH request. This will result in significant reduction in the collision rate in the RACH. A slotted access scheme, where each device will be able to transmit according to its needs will further benefit the system.

According to the proposed solution:

- The devices are being grouped by the network based on their mobility and their communication characteristics (e.g., data to be transmitted, packet delay requirements)
- The network schedules the cluster heads' transmission opportunities based on their transmission requirements. The scheduling information includes every how many timeslots each device should attempt to access the network and also which preambles should be used.
- The intra cluster communication may take place either via a different interface or via scheduled D2D communication.

Compared to the SOTA, which applies device grouping based on the mobility and the positioning we propose that the devices with stringent delay requirements shall have priority in transmitting. Such devices should be selected for aggregating requests from other devices with relaxed delay requirements. The latter decision implies that on the one hand reduced number of devices will compete for RACH resources thus resulting in less collisions and on the other hand that since these devices have stringent delay requirements, if successful they will ensure that the delay requirements of the devices with relaxed delay requirements will be satisfied as well.

Figure 5-2 presents the message exchange for the grouping. During the initial attach message, the UE will provide to the grouping function located in the MME the UE network capability, avg.

RSRQ per cell ID, location. Once the grouping function (in this message exchange is coupled with the MME) will obtain the respective service information from the HSS for this UE; the latter include service access periodicity, delay tolerance, mobility status, etc. Based on these information fields the UE will perform the devices grouping ensuring that the device delay requirements can be covered by the UE that will aggregate the requests; at the same time the mobility patterns of the UEs are taken into consideration for ensuring the cluster stability. In other words, UEs that have the similar or the same mobility pattern will be grouped together for avoiding very often regroupings. The devices that have the more stringent delay requirements are selected as cluster heads for ensuring that they will have a priority in attempting to access the system. If two or more devices have the same delay requirements, the radio measurements are considered.

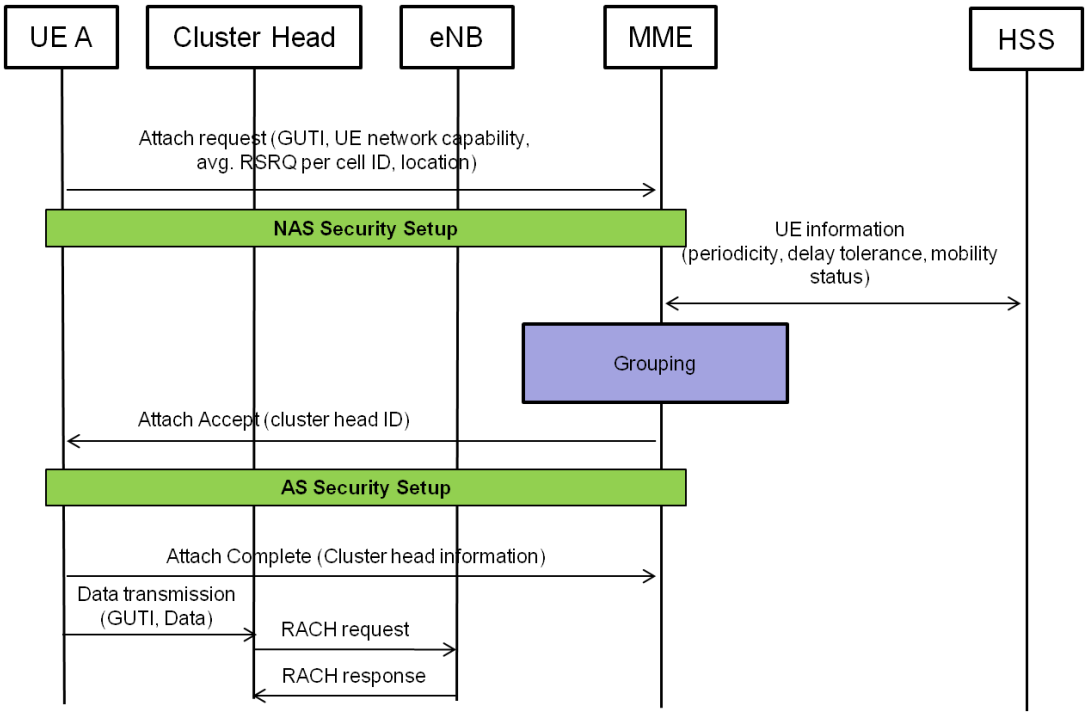


Figure 5-2: Signalling exchange for grouping and for RACH attempt

As Figure 5-2 illustrates, once a UE is grouped by the network, this information is transmitted back to it and the UE associates to the cluster head that the network has indicated. Every time that the UE has to transmit data the cluster head either aggregates only the RACH request or the data as well and performs a joint transmission The second approach is illustrated in Figure 5-2.

In the evaluation, a single BS deployment is considered and the devices that are trying to access the network are static or semi-static. The intra-cluster communication is considered to be either scheduled D2D or it is performed via Uu interface. The devices that are accessing the medium may have either periodic or totally random transmission attempts.

In the simulation analysis the devices are accessing the system simultaneously. In case a collision occurs the devices proceed in retransmissions considering the service requirements (depending

on the urgency of the accessed service more retransmissions are allowed). The devices for their initial access are selecting randomly one out of the 64 available preambles and the retransmission process follows LTE-A approach. The periodic system access and the limited mobility of the devices enable the allocation of devices to groups. The groups are stable since a small number of changes are foreseen for the mMTC devices. Table 5-1 provides the simulation parameters of the group based mechanism compared to LTE rel. 12.

Table 5-1 Simulation Setup

Simulation Setup	
Number of devices	[10.000, 100.000]
Number of eNBs	1
Random Access Opportunities	1 per time frame with 64 preambles
Transmission periodicity	Periodic (1/minute, 1/hour, 1/day) or asynchronous
Data Transmission size	20, 75 or 125 bytes
Acceptable packet delay	Small delay [5,10] msec or Larger delay [1,5] sec
Devices' mobility	Stationary or Low mobility (i.e. < 3Km/h)
Simulation area	387m x 552m grid area
Devices' distribution	Uniform in the simulation area

From the evaluation results depicted in Figure 5-3 it is observed that using the group based system access reduces the collision rate significantly. Additionally, the average initial access delay (i.e., random access, random access response, terminal identification, and contention resolution) is reduced, as shown in Figure 5-4, since the devices are accessing the system with fewer collisions and thus experience fewer retransmissions. For low number of devices, the collision rate is small since the preambles are enough for the random access. As the number of devices increases the collision rate and the CP latency increases for both approaches, but in the case of the group based access it is considerably lower. This is related to the reduction of the number of the devices that compete for the RACH resources (only the group heads) which reduces the collisions and the consequent delays.

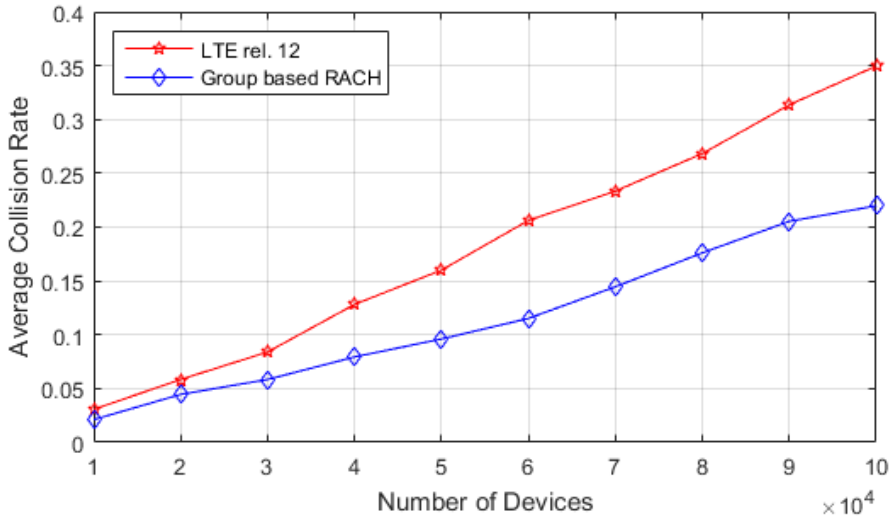


Figure 5-3: Number of collisions for the Group Based System Access compared with LTE-A.

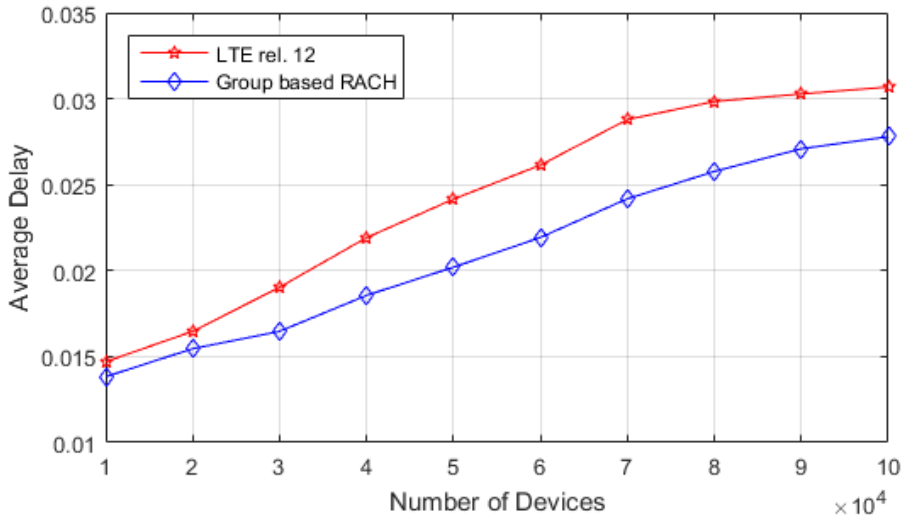


Figure 5-4: Average initial access delay of successful system accesses for the Group Based System Access compared with LTE-A

Figure 5-5 and Figure 5-6 presents the collision rate and the average delay for the high priority devices, which are prioritized; additionally these devices they do not wait for performing their retransmissions in case of collision. We observe that for the high priority devices in large numbers of competing devices we have significant gains compared to the LTE. However, given the fact that the devices are not attempting to access the system again after a failed set of retransmissions, the collision rate is rather high. For the delay we observe that the proposed scheme outperforms the LTE rel. 12 by ~20% less delay in the high priority requests.

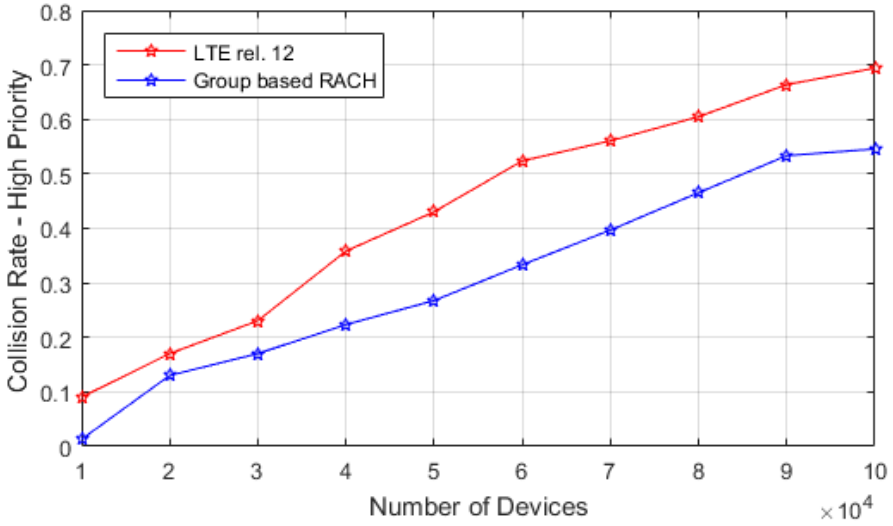


Figure 5-5: Number of collisions for the Group Based System Access compared with LTE-A for high priority requests.

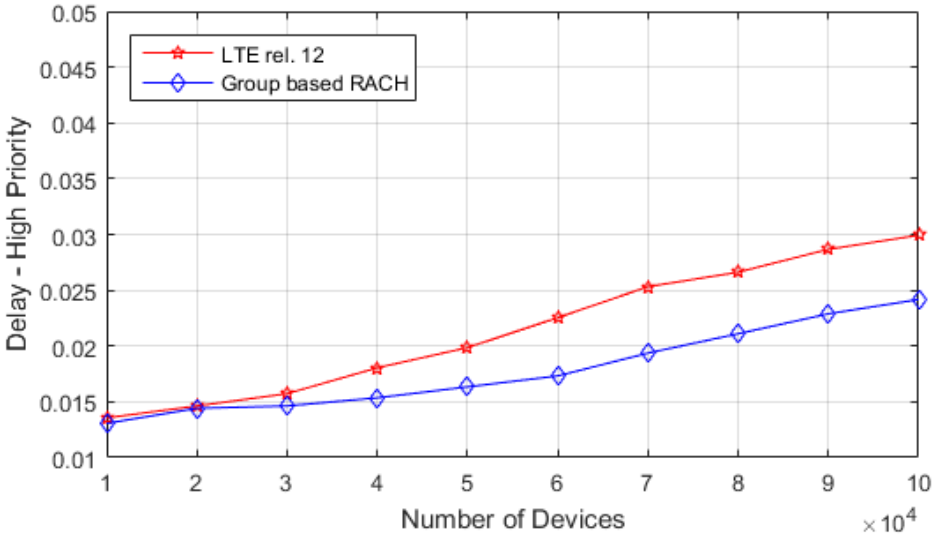


Figure 5-6: Delay for the Group Based System Access compared with LTE-A for high priority requests

5.1.2 RACH Multiplexing in Support to Diverse Access Requirements

One possible solution for the devices with strict latency requirements could be to reserve a set of dedicated preambles for the use of devices with high priority. This solution however is not efficient, since the number of RACH preambles is very small (i.e., 64 preambles) which has to be used both for random access and for handover purposes.

In order to provide an efficient prioritization mechanism for delay-sensitive services (not relying on the assignment of dedicated preambles) METIS-II is currently investigating random access solutions to provide some level of access differentiation per service, taking their accessibility requirements into account.

In the currently proposed solution, random access requests associated with delay sensitive services could be configured to apply a combination of preamble signatures at a given random access time slot. The aforementioned approach would enable requests with stricter delay requirements to have higher priority, since combinations of preambles can always be identified by the receiver. This way, requests with higher priority are significantly less prone from collisions and the retransmissions (Figure 5-7).

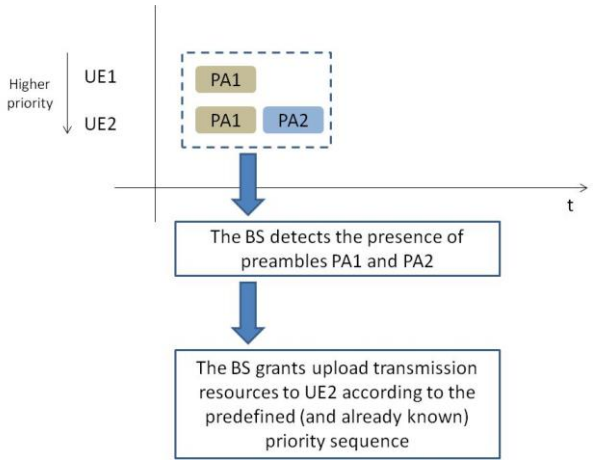


Figure 5-7: Preamble combination for prioritized UE

As it is shown in Figure 5-7, the prioritized UE uses a combination of the preamble signatures at one random access time slot to “overwrite” the other preambles. More specifically, UE2 is high priority compared to UE1, thus it sends a combination of the preambles. Preamble PA1 (2 times) and PA2 can be well detected at the RAN receiver respectively. Hence, the receiver detects the preambles PA1 and PA2 and identifies this combination as a high priority request. The proposed solution guarantees the priority of a particular request in the random access procedure. The high priority request doesn’t need to enter the back-off and retransmission procedure in case of the collision, so that the delay caused by collision is minimized for the high-priority request.

Evaluation of this approach has been done for METIS-II Use Case 4. The devices were abstracted to the number of random access attempts per second ranging from 10,000 to 50,000. 10% of these attempts were associated with delay sensitive services (e.g. uMTC) and remaining ones were caused by services with no strict delay requirements. Three approaches for use of RACH preambles were analysed. In the first one, all devices (high and low priority ones) are accessing the system using one preamble (denoted in the figures as LTE with full resource pool). In the second one, RACH requests coming from delay sensitive services (high priority transmissions) were given a fixed number of 20 out of 50 available preambles (we assume that a number of preambles is used for contention free RACH), while remaining 30 were shared between delay-

tolerant applications (low priority transmissions); this is denoted as LTE with resource split. In the third approach the proposed scheme is applied (denoted as Preamble Coding), where two preambles are used for high priority initial access and one for low priority initial access.

As depicted in Figure 5-8 (a) for high priority initial access assessment, the preamble coding outperforms the other two approaches, since the combination of preambles reduces the probability of having a collision when accessing the system. Collisions in high priority transmissions will occur only if two low priority devices select the same preambles as the combination dedicated for a high priority device that attempts to access the system at the same time. As it can be seen the proposed scheme for the high priority requests the proposed system has ~1000 times lower retransmission probability compared to the LTE without resource split in large numbers of random access rates whereas it has ~100 times lower retransmission probability compared to the LTE with resource split

For the low priority requests the LTE without resource split scheme outperforms both the proposed scheme and the LTE with resource split scheme, as shown in Figure 5-8 (b) since there is no differentiation among the service requests and all the preambles are used for the low priority requests as well. It should be noted though that the preamble coding scheme performs significantly better than the LTE with resource split scheme because all the available preambles are used.

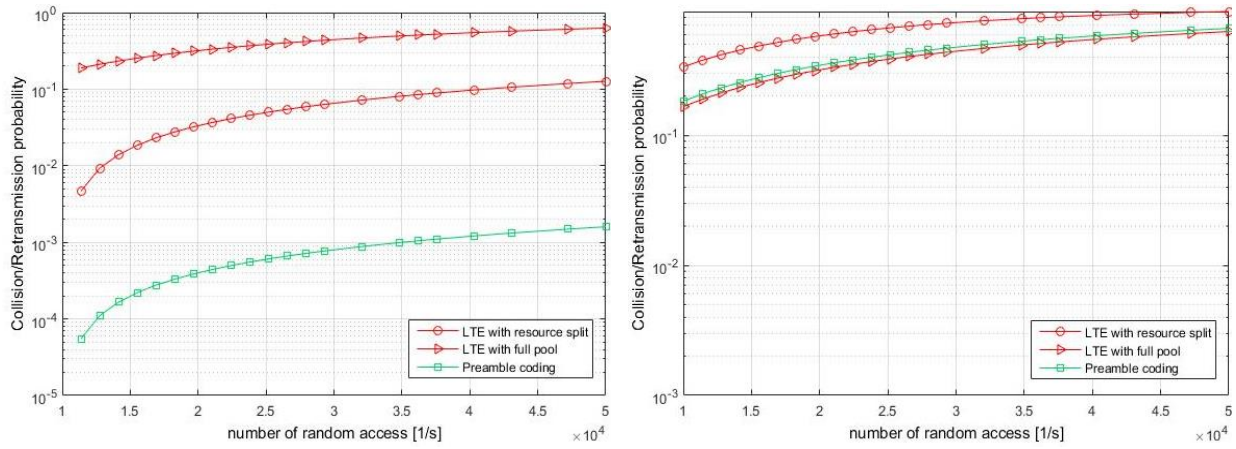


Figure 5-8: (a) Comparison of collision/retransmission probability for high priority request. (b) Comparison of collision/retransmission probability for low priority request

5.1.3 Synchronization Sequences and Random Access in Higher Frequencies

Beamforming is expected to further increase the challenges of initial access in general [MII16-D61]. In LTE-A omnidirectional transmission is considered for synchronization signal(s) and system information i.e. there is no directionality and the channel is more robust. However, considering that new AIVs should be possibly deployed in higher frequencies more quick drops of SINR are expected due to channel intermittency and beams are expected to cover smaller

areas and beam to beam handovers will be more frequent. Thus, efficient synchronization and system access procedures are needed for handling the more frequent handovers and cell (re-)selections.

The major difference of the designs presented and evaluated in D6.1 [MII16-D61] compared to the current LTE initial access procedures, are that the two initial steps, Sync and RA, are done directionally, and Sync and RA preambles are spread across four different frequency regions to gain frequency diversity.

The key outcome of the analysis is that digital BF in the initial access phase (and control messaging in general) offers tremendous gains in terms of latency (consider fast transition from idle to connected state) and overhead, and it should therefore be seriously considered for a mmWave AI design. However, the last statement stands provided that the beam search has already been completed. Otherwise, employing analog BF would bring a huge burden on the system in terms of delay. This delay not only is larger than when digital BF is used, but it is also unacceptably larger than current 4G standards.

5.2 Paging

5.2.1 RAN controlled location tracking and initiated paging

The RRC Connected Inactive state assumes that the connection between RAN and Core is maintained during the low activity periods. Therefore, the RAN can control the UE location tracking and paging during the RRC Connected Inactive state which can be done by the anchor gNB that stores the UE context and terminates the NG connection for the UE.

For RAN controlled location tracking, the RAN needs to be partitioned into group of gNBs and cells which are here called RAN tracking areas (RTA). The network needs a configuration of RTAs across the whole network and every cell broadcasts its RTA identity (RTA ID). The anchor gNB provides the UE with the list of RTA IDs that the UE may move without updating its location. If the UE moves out of its list of RTA IDs, it sends a location update to the RAN which may trigger an anchor gNB relocation.

On the other hand, when the anchor gNB receives an MT (Mobile Terminated) data, it triggers paging to reach the UE, see Figure 5-9:. The UE is paged through all the cells in its list of RTA IDs. In case the list of RTA IDs of the UE includes multiple gNBs, a horizontal Paging inter-gNBs interface is necessary. This requires anchor gNB to maintain the inter-gNB relationships with all gNBs of any RTA which it has given to the UEs. In addition, the anchor gNB needs to buffer and forward the UE MT data until the anchor gNB is relocated. Upon receiving the paging message, the UE responds to the paging and is ready to receive downlink user plane or control plane data with existing RRC configuration.

This approach is explained in more detail in [MII16-D61] as part of Hierarchical Paging concept, where it was identified that one of the advantages of RAN controlled location tracking and paging is the potential to significantly reduce the paging load on the air interface using a smaller paging

areas, especially in the case of (semi-)stationary UEs. The location updates in RAN would not be so frequent and the network can use the UE location to proactively forward data to the gNB where a UE is known to be camping.

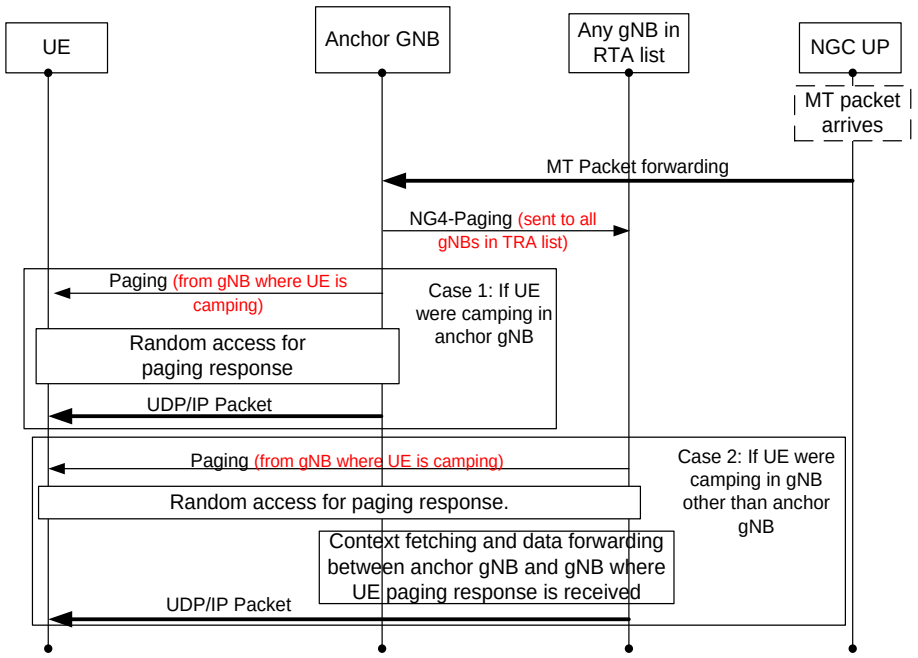


Figure 5-9: RAN initiated paging in the RRC Connected Inactive

5.2.2 Hybrid CN/RAN controlled location tracking and CN initiated paging

An approach where the location tracking and paging initiator can be either in RAN or in Core Network is called hybrid CN/RAN controlled location tracking, which is using the principles of core network based paging (known from LTE) and the hierarchical paging [MII16-D61]. The hybrid paging scheme was proposed to 3GPP [3GPP16-R2167708]. In this approach, the Next Generation Core (NGC) acting in the paging initiator role becomes aware that the UE is in the RRC Connected Inactive state. RAN can decide that the paging of the UE needs to be initiated from the CN due to different reasons. The decision to use CN initiated paging may be made when the UE moves out of its list of RTA or during transition from RRC Connected to the RRC_ Connected Inactive for a fast-moving UE to avoid frequent RTA updates. In general, if the UE moves out of its list of RTA IDs, it sends a location update to RAN which may trigger the anchor gNB relocation.

When an MT data arrives, the NGC user plane (UP) sends a notification message to the NGC control plane (CP). The NGC CP then sends paging to all the gNBs in the list of RTAs of the UE, see Figure 5-10: for paging the UE over the air. Upon receiving the paging message, the UE

proceeds with random access procedure to respond to the paging. If the UE initiates the paging response at a gNB other than the anchor gNB, the gNB that received the UE paging response may request the NGC UP to forward the UE MT data. The NGC-UP forwards the UE MT data to the gNB that received the UE paging response.

Whether the paging is initiated from the RAN or the CN is transparent to the UE if the UE ID included in the paging is the same, or alternative the UE may monitor two different identifiers. In both cases the UE responds to the paging message in the same way.

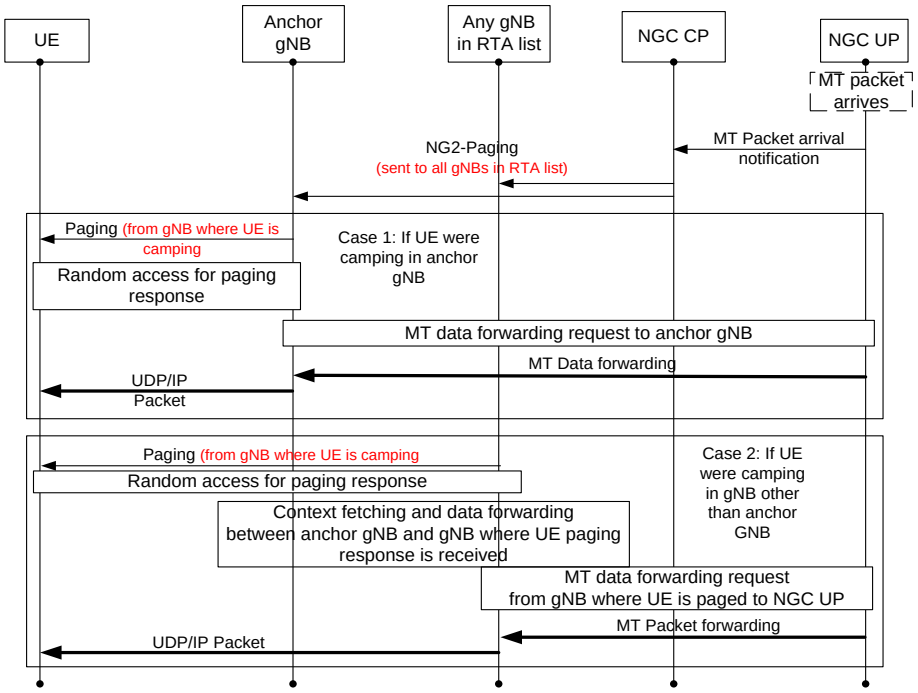


Figure 5-10: Potential CN initiated paging in the RRC Connected Inactive

5.2.3 Performance comparison

For the performance comparison, both described approaches can be evaluated simply by counting the number of messages for signaling overhead and latency assuming the following 5G architecture [3GPP16-23799].

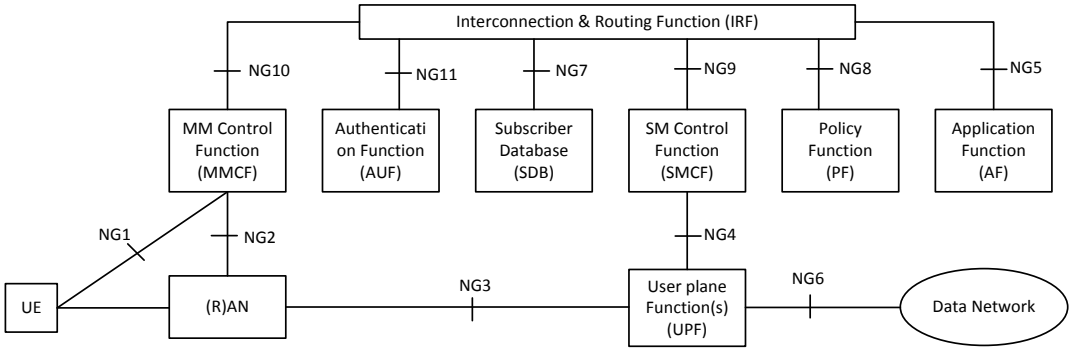


Figure 5-11: Reference architecture for 5G

Based on the Figure 5-9 and Figure 5-10, we can roughly estimate the comparable signaling load in CN and RAN based paging approaches. Let M and N denote the number of cells and gNBs per UE's list of RTAs, respectively. The paging load in the RAN initiated paging is M messages (*over radio*) + (N-1) messages (*over NG4*). It is M messages (*over radio*) + (N+3) messages (*over NG2*) + 3 message (*over NG6*) in the CN initiated paging. We have also added the two NG2 and two NG6 messages that are required to make the NGC CP and NGC UP be aware that the UE is in the RRC Connected Inactive. The extra one NG2 and one NG6 messages are the data notification message sent from NGC UP to NGC CP and the MT packet forwarding request message sent from gNB that receives the UE paging response to NGC UP, respectively. Thus, if M = 1 and N = 1, i.e. the paging area is restricted to a single cell, the paging load in the RAN initiated paging is 87.5% less than the paging load in the CN initiated paging (1 message vs. 8 messages). However, if let say M = 3*19 and N = 19, the percentage of paging signaling reduction is only 8.5 % (75 vs. 82 messages). This indicates, the paging load reduction from the RAN initiated paging is significant when the paging area is small.

The RAN initiated paging also requires data buffering in the anchor gNB and data forwarding across NG4 interface during MT transaction and when the paging is responded to a gNB other than the anchor gNB. This would be a significant drawback if the UE is moving in a network with restricted NG4 interfaces with star topology in transport network. On the other hand, the RAN initiated paging has lower latency than the NGC CP initiated paging. Here, we are referring to the time interval between the arrival of the first MT packet in the NGC UP until it is received by the UE.

Both the RAN controlled and hybrid CN/RAN controlled location tracking could be supported for UEs in RRC Connected Inactive state so that the paging can be done in transparent manner. The benefits for RAN based paging can be utilized for semi-stationary UEs and the hybrid RAN/CN based paging can be utilized for fast moving users and providing load balancing options for network operator.

5.3 5G RAN Lean Design

We often underestimate the importance of NW energy consumption when we only express it in terms of kWh and translate that to operational cost savings or CO2 emission reductions. In addition to the economic and environmental reasons to minimize energy consumption there is also a large number of arguments related to engineering. Today the main obstacle for miniaturization of radio base stations is heat. Energy consumption directly drives product weight and volume through heat dissipation. Reducing energy consumption will have non-linear impacts on not only weight and volume but also on what a base station is, how we build it, what it can do, where we can deploy it, what energy sources we can utilize, etc. When the network gets increasingly densified as the traffic demands grows, more nodes are required. If there is no (or very limited) possibility for the base stations for micro sleep (cell DTX) the power consumption may increase when the cell density is increased in the network [ERI11]. One of the drawbacks with LTE was the rather low possibility for the cell to enter a so called micro sleep (the cell DTX). There are several reasons for this related to the system information distribution:

- Reference symbols transmitted even if no data
- System information transmission transmitted periodically regardless of load
- PDCCH, is transmitted across the full system bandwidth i.e. same number of PDCCH symbols are used for all RBs
- Synchronization signals

This section describes ways to make 5G energy efficient than 4G, i.e. **lean design**. The system information distribution is explained in more detailed in Section 5.3.1, while the overhead of system information using LTE is analyzed in Section 5.3.2. Finally, Section 5.3.3 analyses the possible energy efficiency gains of 5G over LTE.

5.3.1 System information distribution

Reference Signal in general

The reference symbols (RS) necessary for channel estimation should in 5G typically only be transmitted in the same subframe, over the same bandwidth, and in the same beam as the corresponding data. This is different from LTE which can also the cell-specific reference signals (CRS) in previous subframes to aid channel estimation. Exactly how this will be done for NR is now up for discussion in 3GPP [3GPP17-38912], [3GPP17-38804].

System information transmission using user on-demand approach

In the 3GPP NR discussions, the system information is divided into minimum SI and other SI. Minimum SI is periodically broadcast (as in LTE today). The minimum SI comprises basic information required for initial access to a cell and information for acquiring any other SI broadcast periodically (as in LTE) or provisioned via on-demand basis (new compared to LTE). The other SI encompasses everything not broadcast in the minimum SI. The other SI may either be

broadcast, or provisioned in a dedicated manner, either triggered by the network or upon request from the UE [3GPP16-R2168858]

PDCCH

In LTE, PDCCH is transmitted across the full system bandwidth i.e. same number of PDCCH symbols are used for all RBs. This is not especially resource and energy efficient. For 5G, we foresee a more efficient PDCCH transmission, probably it will be more limited to the resource used by the user data.

Synchronization signals

LTE uses a periodicity of 5 ms. However, if the periods between the synchronization signals can be increased, the base station sleep efficiency can be increased [DDL15]. The reason is that it takes some time to deactivate and reactivate certain components, and given this the longer the sleep duration, the more components can be put to sleep and the lower the sleep power usage becomes.

5.3.2 System information distribution overhead with beamforming

Information transmission presented afore is considered a key aspect. First of all, it is important to mention that this information is constantly broadcasted, but depending on the type of information, different periodicities are assumed. In LTE the time-domain scheduling of the MIB and SIB1 messages is fixed with periodicities of 40 ms and 80 ms. Furthermore, for the MIB the transmission is repeated four times during each period, i.e., once every 10 ms. SIB1 is also repeated four times within its period, i.e., every 20 ms, but with different redundancy version for each transmission.

MIBs, SIBs, and CRS drive most of the energy consumption in the network regardless of the amount of traffic. In other words, due to the way system access functionalities have been designed (relying on these signals), the energy consumption in mobile networks do not scale with the amount of consumed traffic i.e. the User Plane (UP) transmissions. In other words, the way these signals are transmitted cannot be configured based on the traffic scenarios e.g. in high traffic areas or low traffic areas.

The analysis presented in [MII16-D61] concluded in that the overheads for MIB, SIB1, and SIB2, are significant and range from 4-8% for each one. Consequently, the total overhead is in the range of 11-15 % (assuming beamforming, see details of the assumptions in the same document). Nevertheless, it can be seen that the overhead from access information broadcast has the potential of consuming a significant part of the system capacity.

5.3.3 Energy efficiency evaluation

For evaluating how a lean system can reduce the energy consumption an energy efficiency evaluation is done for different cell densities, going from 1000 to 100 m with a fixed number of users in the area, i.e. the number of users per cell decreases with *decreased* cell radius. The

model description, assumptions and parameters can be found in Annex A.1, here we just show a summary of the results. Figure 5-12 shows the relative power consumption per cell for 5G compared to LTE. The major difference between 5G and LTE is the ability to utilize the Cell DTX, see Table A-3 in Annex A.2. In Figure 5-12 LTE has a Cell DTX probability of 16% [TFA+16] and NR 71% [TFA+16] and 90%. As can be seen, the power consumption is lower for the 5G case due to 5Gs improved ability for cell DTX compared to LTE. For example, the decrease in power consumption for 5G at 200 m is around 50% when there are 5 users/km^2, 30% when there are 20 users/km^2 and 15% when there are 50 users/km^2. In general, the power savings is in the same ballpark as in [MII17-D23] (Section 3.3) and [TFA+16].

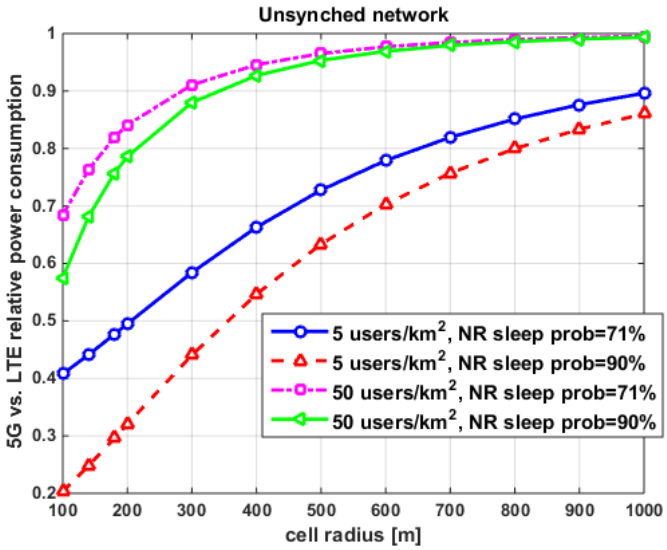


Figure 5-12 The 5G power consumption per area [W/km^2] when the system is densified vs. LTE for different NR Cell DTX probabilities. LTE has a 16% Cell DTX probability. Note that there are 20 users per km^2 regardless of cell radius which means that the probability for zero users per cell increases with *decreased* cell radius.

This results show that if the 5G is design so that it allows better Cell DTX sleep probabilities than LTE, the power consumption can be decreased substantially.

6 Mobility

6.1 Introduction

This chapter describes the specific mobility related challenges to the design of components necessary for the design of the mobility procedures. Mobility in the 5G framework needs to cover use cases with active users in RRC Connected state and low activity users in RRC Connected Inactive state and in RRC Idle state. In addition, this chapter presents new ideas on how to explore the usage of tight interworking between LTE and 5G (NR) for mobility users. Finally, the UE perspective is given for mobility challenges including UE capability signaling and LTE-5G tight-interworking scenarios from UE perspective.

The key take-away in this chapter presents the 5G mobility framework consisting of several new methods including UE autonomous mobility, make-before-break handover and mobility concepts for URLLC. Make-Before-Break is thoroughly investigated in this deliverable together with considerations of multi-connectivity for a centralized RAN architecture. Centralization of the UE AS context, RRM, and multi-connectivity have potential to improve the efficiency of various Connected Inactive and Connected state mobility and multi-connectivity procedures. Inter-RAT mobility between LTE and 5G allows seamless inter-RAT mobility and multi-connectivity by anchoring the RAN/CN interface to the LTE or 5G network node. Inter-RAT mobility can support UEs in both RRC Light Connected in LTE and RRC Connected Inactive in 5G state thus allowing low energy dissipation and fast state transition to Connected state from either of the systems in inter-working scenarios. Finally, UE context information is used to enhance the accuracy of mobility prediction enabling a uniform service experience for the user, even in deep shadow regions or coverage holes. User profiles are used for predicting the future network requirements. This approach reduces the signaling cost for context information transfer.

6.2 Mobility and Multi Connectivity in centralized RAN architecture

6.2.1 Problem description

The path towards 5G will comprise a common core to support multiple access networks that are transparent to the end-user. Multi-connectivity and aggregation solutions will therefore be key ingredients of the 5G RAN architecture. A wide variety of services stemming from E-MBB, M-MTC, and URLLC means that networks must also offer different and tailored performance capabilities. Furthermore, 5G shall be able to exploit the existing network assets in a flexible and cost-effective manner.

One potential architecture meeting these goals comprises a multi-layer radio access network [NOK16] where the baseband processing architecture is divided into real time (RT) and non-real time (NRT) functions. The non-real-time baseband comprises PDCP, RRC, and possibly the non-

real-time functions of the RLC protocol, such as ARQ. The real-time baseband comprises the lower layer protocols, such as physical layer, MAC, and RLC. Such architecture also supports the evolution of legacy cloud RAN architectures, which are defined mainly by the transport capabilities. The multi-layer RAN is illustrated in Figure 6-1 below:

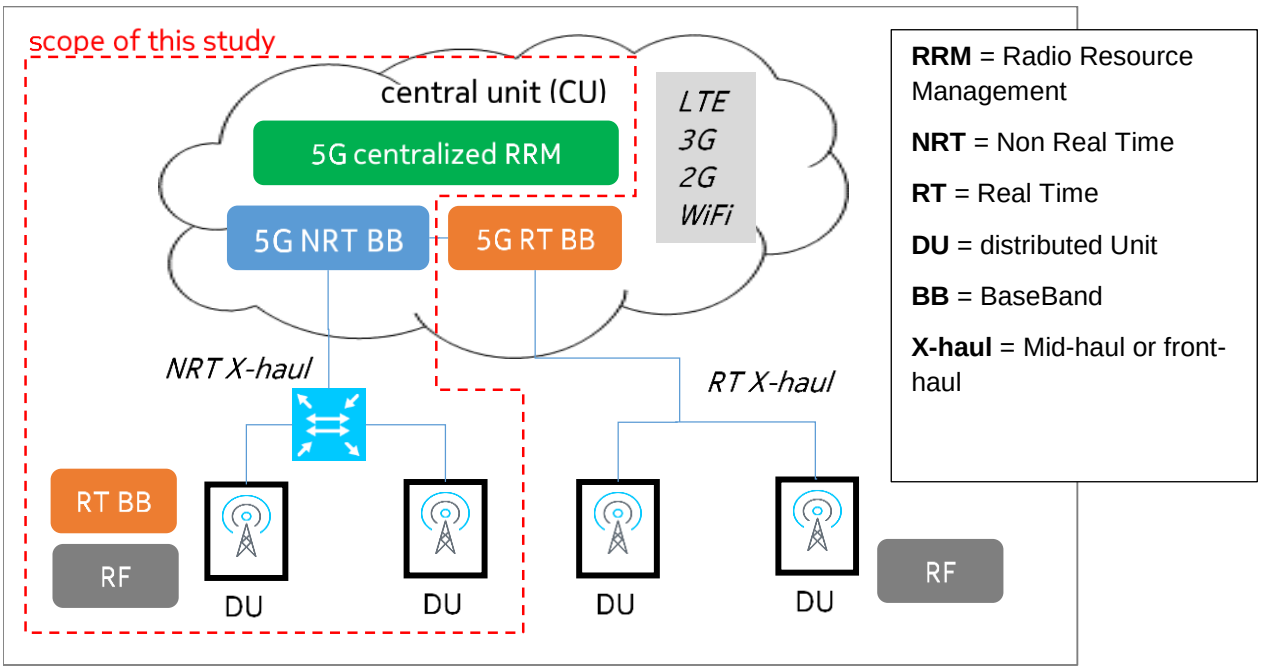


Figure 6-1 Multi-layer RAN

The central unit (CU) hosts the NRT baseband, centralized RRM functions, and RT baseband functions for DUs connected with a RT X-haul (mid-haul or front-haul depending on the architecture). It is the main aggregation point for intra-5G and inter-RAT multi-connectivity, allowing virtualization of 5G in the same node with legacy technologies (2G, 3G, WiFi, LTE). The CU may be physically located e.g. in a large data center or a macro base station.

The distributed unit (DU) hosts the Radio Frequency (RF) and RT baseband functions for DUs connected with NRT X-haul. A DU may be physically located e.g. at distributed RF sites, centralized local aggregation hubs, or at large data centers, hence providing flexibility to optimize the network per latency requirements of the service, or front-haul technology. The RT baseband may be co-located with the RF site or connected by a fast (e.g. fiber based) front-haul interface. The NRT and RT functions are connected by a 'non-ideal' mid-haul interface (e.g. Ethernet) with relaxed requirements for latency and throughput.

The terms "centralized architecture", and "CU/DU architecture", are interchangeably used in this study to refer to an architecture where the NRT functions reside in CU, while the RT functions reside in DUs. The RF functions, and possibly a part of PHY, may be co-located with MAC and RLC (one tier architecture) or reside in remote units (two tier architecture).

Given above definitions, the problem addressed by this section is:

How to exploit centralized RAN architecture to further optimize the 5G mobility and multi-connectivity, addressing key 5G design targets such as:

- extremely short service interruption (URLLC)
- robust UE mobility on beamformed layer (URLLC, MBB)
- low signaling overhead (massive MTC)
- fast resume of connectivity (URLLC, MBB)

6.2.2 INACTIVE state mobility

Intra-frequency mobility

In RAN-controlled location tracking, the access network is partitioned into groups of cells called RAN tracking areas (RTA). The UE is configured with a list of RTAs it may move within without updating its location.

In centralized architecture, the AS context of the RRC_CONNECTED_INACTIVE state UEs is mainly stored in the CU, hence keeping the UE mobility hidden from the core network (avoiding path switch) when UE moves within an area covered by one CU. Furthermore, only part of the UE context needs to be fetched to the new DU when UE moves from a cell covered by one DU to another.

When UE enters a RTA covered by a new CU, the context is fetched to the new CU. Hence the UE context is always in the right RAN node when UE resumes the RRC connection or carries out a small data transmission.

The area where the UE can move without updating its location (RTA list) can range between a single cell and the cells covered by one CU. A small location area minimizes paging at the expense of location updates, while a large area implies the opposite. The area can be dynamically adjusted, for example, per UE speed (location update rate) and state transition rate (paging rate), exploiting the global view of the DUs controlled by a CU.

In downlink based mobility, UE carries out cell selections and re-selections based on downlink reference signal. Alternatively, RRC_CONNECTED_INACTIVE state mobility in a cloud comprising uplink- and downlink time synchronized cells (e.g. small cells under one macro) may be based on a hybrid approach, where the UE location is mostly tracked per UL-RS, however providing DL-RS at the edge cells of the CU area to allow UE to report a change of CU area. Such UL-RS based UE tracking benefits from centralization as the beacons can be configured and the beacon transmissions gathered and processed by one protocol entity i.e. the RRC located in a CU.

Inter-frequency mobility

A typical strategy in a heterogeneous network (HetNet) comprising high capacity small cells and a macro coverage layer would be to prioritize macro layer at least for the mobile INACTIVE state

UEs with no multi-connectivity support. This is to avoid excessive signaling from location updates, paging failures, and problems from camping on a layer exploiting analog or hybrid beamforming. However, when a large amount of data arrives to UE or network buffer, it becomes desirable to re-direct the UE to high capacity small cell layer as fast as possible with minimal amount of extra signaling.

Re-direction during connection establishment is not supported in LTE, but the UE needs to go through a normal establishment procedure followed by a handover to small cell layer. This procedure can be optimized for the RRC_CONNECTED_INACTIVE state, exploiting centralized architecture in re-directing the UE to small cell layer as part of the connection resume procedure. Consequently, the paging, location updates, and small are effectively transmitted via macro cell while larger amounts of user data are delivered via high capacity small cells.

6.2.3 ACTIVE state mobility

The main new challenges of 5G mobility are the strict requirement for service interruption time (down to zero milliseconds from UE point of view for URLLC service) and sufficient robustness in high frequency (HF) bands exploiting analog or hybrid beamforming. Two promising techniques to overcome these challenges, respectively, are the so called make before break handover and UE autonomous handover with early provision of the handover command. These techniques can be applied independently or together to realize a seamless and/or robust handover. In the following, we discuss how they can benefit from centralized architecture.

Besides providing benefits for the new handover paradigms, the centralized architecture can improve the network controlled break before make paradigm that is being used in legacy systems such as LTE. Details can be found in Annex A.7.

Make before break handover

Make before break handover is an essential component in reaching the sub-ms service interruption required by some 5G services. One of the biggest challenges of *intra-frequency* MBB is the need to maintain communication links towards both source and target cell. This is illustrated in Figure 6-2 below, showing parallel transmissions/receptions from the source and target DUs in red and green, respectively.

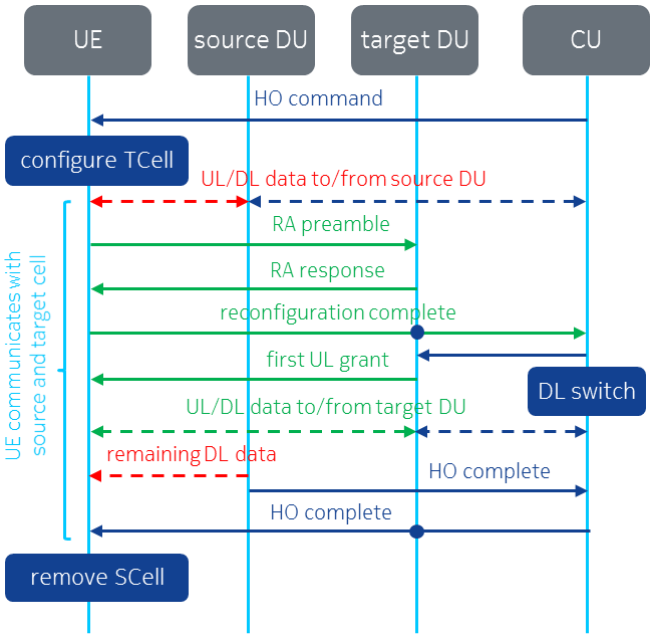


Figure 6-2 Make before break handover in centralized RAN architecture

The MBB HO can be realized by one of the following principal means:

1) Utilizing a duplex gap to separate the uplink and downlink transmissions. This is not applicable for TDD, which is expected to be the main duplex technique for 5G small cell deployments.

2) Applying interference cancellation to mitigate the UL-DL interference (UE self-interference), UL-UL interference (UE to UE interference), and DL-UL interference (BS to BS interference) in TDD based deployments. Efficient mitigation of the UE self-interference is a subject of active research, but may not be mature enough in the expected 5G timeframe.

3) Aligning the UL and DL TDD frames of the source and target cell. This is a feasible for static or semi-static TDD, but non-straightforward for dynamic TDD where the transmission direction of each cell may ultimately change every subframe.

The MBB HO in dynamic TDD deployment (method 3 above) benefits from centralization by the means of more flexible coordination of the transmission direction of the source and target cell. The downlink and uplink frames of the source and target cell can be temporarily fixed in the presence of MBB handover, applying dynamic TDD otherwise. Upon MBB handover, CU configures the source and target DU with a pattern of fixed UL and DL frames, considering also possible MBB handovers to/from other DUs. After completion of the HO, the fixed frames are released for dynamic TDD. Such coordination is more challenging in a distributed architecture, as a gNB has poorer visibility to its neighboring gNBs.

As another benefit of centralization, a change of the master and secondary roles of the source and target cell is not needed. The role change is required in a distributed architecture for relocating the RRC protocol from the source node to target node. In centralized architecture, there is no master-slave relationship between the source and target DU, as both are served by a common CU that contains the RRC protocol.

UE autonomous handover

UE autonomous handover is a potential technique for lowering the probability of handover failures (HOF) in the case of a fast degradation of the communication link to source cell. Such blocking is particularly a problem in high frequency bands, where the SINR may be degraded in the order of 20dB/100ms [3GPP17-R21700122]. The basic procedure for UE autonomous handover is illustrated in Figure 6-3 below, omitting steps after the random access to target cell.

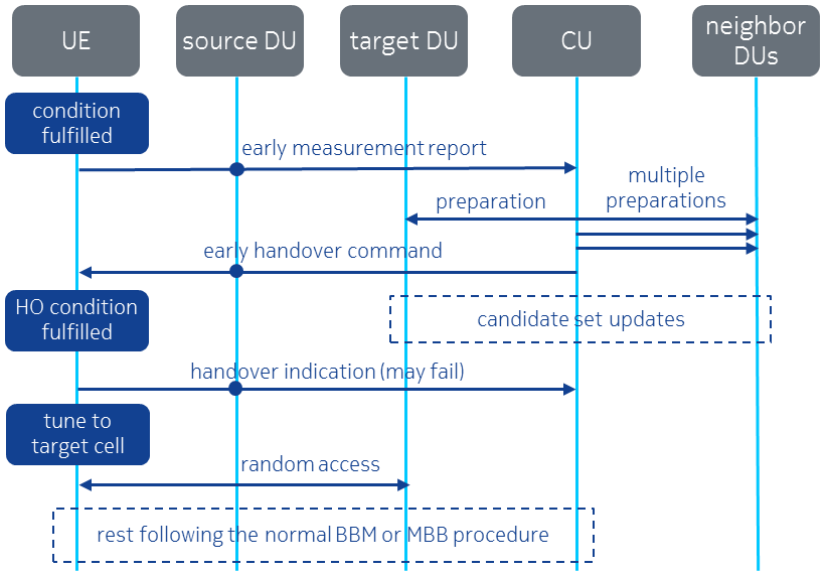


Figure 6-3 UE autonomous handover in centralized RAN architecture

In UE autonomous handover, the UE triggers a measurement report earlier than in the normal handover per condition configured by the network. Based on the measurement report, CU prepares neighboring cells that are the most likely candidates for the handover. CU then sends a handover command to UE, including necessary information for the UE to autonomously carry out the handover to given target cell (e.g. handover trigger, RACH preambles, and SI). Finally, UE evaluates the handover condition based on DL-RS, and autonomously hands over to the target cell when the condition is satisfied.

There are some challenges associated with the UE autonomous handover, which might be alleviated in the centralized architecture:

Preparing candidate cells: Each preparation of a candidate cell involves a signaling exchange between a source and a candidate node. Besides moving straight from one cell to another, the UE may repeatedly cross the conditional handover border in the source cell, each time triggering

a preparation procedure. Furthermore, the UE may move within the "conditional handover corridor" in the cell edge, requiring a mechanism for updating the set of candidate cells. Preparing multiple cells can be hence costly in the distributed architecture in terms of network signaling.

A full context fetch is not needed in CU/DU architecture, but the lower layers of candidate DUs can be prepared for a possible handover. In some cases, the preparation phase might be completely omitted, given CU has all the necessary information (e.g. available RACH preambles, system info) to be included in the early HO command. Admission control is not needed in any case, as the load situation can significantly change during preparation phase.

User-plane handling: In case handover indication fails, the source node does not know when to stop providing UL/DL grants (until radio link failure) and start forwarding downlink packets to the target node. Hence, it would be beneficial if the source node could predict the occurrence of radio handover, minimizing the service interruption due to delayed data forwarding and avoiding interference due to continued downlink transmission. Such prediction can be aided in centralized architecture by a global view of the nodes participating to conditional handover.

Fast activation of multi-connectivity

One challenge of the 5G multi-connectivity is supporting a fast resume of the radio legs from RRC_CONNECTED_INACTIVE state. For single connectivity, a fast resumption of the connectivity can be achieved using a scheme where the network can suspend the connection and store the UE context in RAN. Such resume procedure can be combined with a secondary eNB change to speed up the resume of multi-connectivity. It is assumed that this way the resumption of multi-connectivity is faster compared to full RRC_IDLE to RRC_CONNECTED state transition including activation of multi-connectivity. However, the resumption of multi-connectivity may not be efficient enough for URLLC applications due to significant signaling overhead and associated delay.

The main difficulty comes from the fact the multi-connectivity consists of radio legs which are allocated from different radio nodes. A full resumption of the connectivity in LTE comprises the following steps:

- RRC connection resume including random access to PCell and measurement configuration
- Averaging of Primary Secondary cell (PSCell) measurements and reporting those to network
- SeNB addition
- RRC connection reconfiguration (PSCell addition)
- Random access to PSCell
- Possible data forwarding and path update

Combined, these steps may take up to 1 second, being PSCell measurements the main contributor. Consequently, multi-connectivity cannot be applied to serve occasional bursts of data with high data rate, leading to inefficient utilization of the UE power resources. Furthermore, the

network may not be able to suspend RRC connection in a dynamic manner (e.g. in the time scale of hundreds of milliseconds compared to 5-30 s of LTE) as the multi-connectivity cannot keep up with the state transition dynamics. Consequently, the UE will spend more time in the RRC_CONNECTED state, resulting to worse energy efficiency and increased signaling load.

A multi-connectivity session can be resumed in a fast manner by the help of uplink beacons. In such scheme, the UE is configured to transmit uplink beacon signals in RRC_CONNECTED_INACTIVE state. The beacons are received by DUs and delivered to the CU for further processing. CU maintains a set of candidate links per UE, representing the preferred choices for a multi-connectivity session. In addition, CU could maintain cell-specific uplink timing advance values per UE. Upon arrival of uplink data, UE carries out random access to camped cell and sends *RRCConnectionResumeRequest* message to network as part of the initial uplink message. Upon arrival of uplink or downlink data, network sends RRCConnectionResume message to UE, including a command to resume multi-connectivity to indicated cells. This message may comprise the cell-specific timing advance value(s) to be applied for subsequent UL transmissions. After these steps, the UE is ready to receive and transmit data utilizing multiple communication links.

Detailed signaling procedure can be found in Annex A.8.

Besides faster activation of the full data transmission capability, beacon aided activation implies a lower UE power consumption due to higher utilization of frequency domain as opposed to time domain, and lower signaling overhead per activation. It greatly benefits from centralization, as one network node (CU) has the control of configuring beacon transmissions and processing beacons from multiple radio nodes (DU).

6.2.4 Benefits of proposed concept

The benefits of the centralized RAN, relevant to 5G mobility and multi-connectivity, can be summarized as:

- Anchoring the RAN-CN interface to a common RAN node (CU).
- Anchoring the UE AS context to a common RAN node (CU).
- Providing a global view and control of the radio nodes (DU) for intra-CU mobility and multi-connectivity.
- Facilitating UL-RS based mobility.

These benefits can be exploited by various INACTIVE and ACTIVE state procedures, reducing signaling overhead, decreasing handover interruption time, providing faster activation of the multi-connectivity, and reducing the UE power consumption, as described above.

As a drawback, centralized RAN requires an additional interface that contributes to the overall radio network delay. Hence centralization may not be a preferred choice for services with extremely tight requirements for access delay. This challenge may be solved by a flexible

mid/front-haul architecture [3GPP16-R3162728] which adapts to the deployment scenario and service in hand.

6.3 Inter-RAT mobility

6.3.1 Introduction

The 5G network is envisioned to comprise multiple Radio Access Technologies (RATs) that may operate on different air interface variants [MII16-D61]. This requires interworking between the RATs during inter-RAT mobility. The level of interworking depends on the service requirements service level agreements and capability of a 5G device. For example, a mobile broadband device requires a high data rate, which can be achieved by aggregating resources from multiple RATs. In such cases, a tight interworking between, e.g. Enhanced Long Term Evolution (eLTE) and 5G New Radio (NR), would be crucial to create multi-connectivity to both RATs in order to achieve high data rate for the UE. Other services may have strict interruption time requirement during handover that it is difficult to fulfil using the legacy inter-RAT handover approaches. Such scenarios would also require tight inter-RAT interworking in order to minimize the interruption time during inter-RAT handover. Inter-RAT interworking during inactive state mobility is also required due to state transition latency requirement for some of the 5G use cases. The focus of this section is the interworking between eLTE and NR during inter-RAT mobility. This is considered for both inactive and active state mobility.

6.3.2 LTE-NR Mobility Alternatives

In Table 6-1 the characteristics of different mobility alternatives are shown. The Mobility robustness is increased with increasing complexity of UE (and for the base station). Service continuity means that the bearer and the QoS can be retained even though there might be a transmission interruption. The higher capacity for the Fast user plane switch, DC and CA alternatives comes from the possibility to do load balancing and to some extent the possibility to select the best AIV on a fast basis (see [MII17-D52]).

Table 6-1 Characteristics of different Mobility alternatives

Mobility alternative	Mobility robustness	Data-rate aggregation	Higher capacity	Minimum UE complexity
Cell reselection	Service continuity data	No	No	Single Rx-Tx
Hard HO CN	Service continuity voice/data	No	No	Single Rx-Tx
Hard HO RAN	Service continuity voice/data	No	Slow load balancing	Single Rx-Tx
Fast user plane switch	Quick recovery	No	Yes, fast load balancing / AIV selection	Single Rx-Tx

DC full (DL and UL)	Seamless (0ms interruption)	DL/UL	Yes, Fast load balancing / AIV aggregation	Multi Rx-Tx
CA	Seamless (0ms interruption)	DL only	Yes, fastest in DL	Multi-Rx
TRP level mobility e.g. CoMP	Seamless (0ms interruption) at least for intra-CU mobility	No	Yes, especially at cell edge	Single Rx-Tx

METIS-II view is that LTE-5G tight integration using PDCP layer as aggregation layer is the most feasible solution. There are several reasons for this, one is that making a CA solution between LTE and NR will be rather complicated and may hamper NR to evolve. This is also the case in 3GPP, where the CA alternative for LTE-NR tight integration was ruled out [3GPP17-38804]. However, note that CA with NR will still be studied.

The disadvantage of DC (and CA) is that the UE must be able to communicate with more than one BS at the same time, i.e., dual radios must be supported. For Fast user plane switch, it is enough to have one transmitter (Tx) and one receiver (Rx) since the UE probably can keep both control links by means of time multiplexing operations to listen/measure one RAT at a time.

For the Fast user plane switch alternative, the control plane is connected to both AIVs at the same time but the user plane is only transmitted via one of the AIVs, i.e. the PDCP level routes the user plane packets to one of the AIV nodes. If the control plane is connected to both the LTE node and the 5G (NR) node, no signaling is required to switch and the user plane switch may be almost instantaneous. The fast user plane switch can be based on normal handover measurements such as RSRP, but also more advanced and faster type of measurements are advantageous. Note that for the Fast user plane switch the UE must still be synchronized to the SeNB and receive and send signaling data such as system information, transmit measurement reports to the SeNB, etc. It is expected that the UE can still use a single RX and TX by using an internal time multiplexing operation.

6.3.3 Inter-RAT mobility during inactive state

The inter-RAT interworking in legacy systems during inactive state mobility is based on cell reselection. When a UE reselects a cell in another RAT, it registers to and enters RRC Idle state in the new RAT. Similar approach can also be considered for inter eLTE/NR mobility during inactive state. While that it is relatively simple to implement, it might not be sufficient to fulfil the strict requirements, for some of the 5G use cases. Also, it would be more efficient to utilize the characteristic similarity of the low activity states in eLTE, e.g. Light Connected and 5G NR, e.g. RRC Connected Inactive.

Autonomous inter-RAT cell reselection with context fetching

In this approach, the UE is given a RAN notification area that consists of cells that belong to eLTE and NR. Thus, the UE can autonomously reselect cells between the RANs within its RAN

notification area during the low activity periods. During a Mobile Terminated (MT) call, the network may page the UE through the cells in its RAN notification area. This requires a support for inter-RAT paging. If the UE needs to send Mobile Originated (MO) data, it sends Resource Request message to the cell that is camping on. If the cell is located in a new RAT, the NB fetches the UE context from the anchor NB that is located in another RAT, see details in Figure 6-4:a.

One of the advantages of this approach is that it allows the UE to benefit from the advantages of Light Connected of eLTE and RRC Connected Inactive of NR. It leads also to a less UE power consumption and signaling overhead from inter-RAT mobility when compared to the legacy approach. However, the approach might be difficult to implement considering the UE need to autonomously change states between Light Connected and RRC Inactive, especially if the states don't have perfectly aligned characteristics.

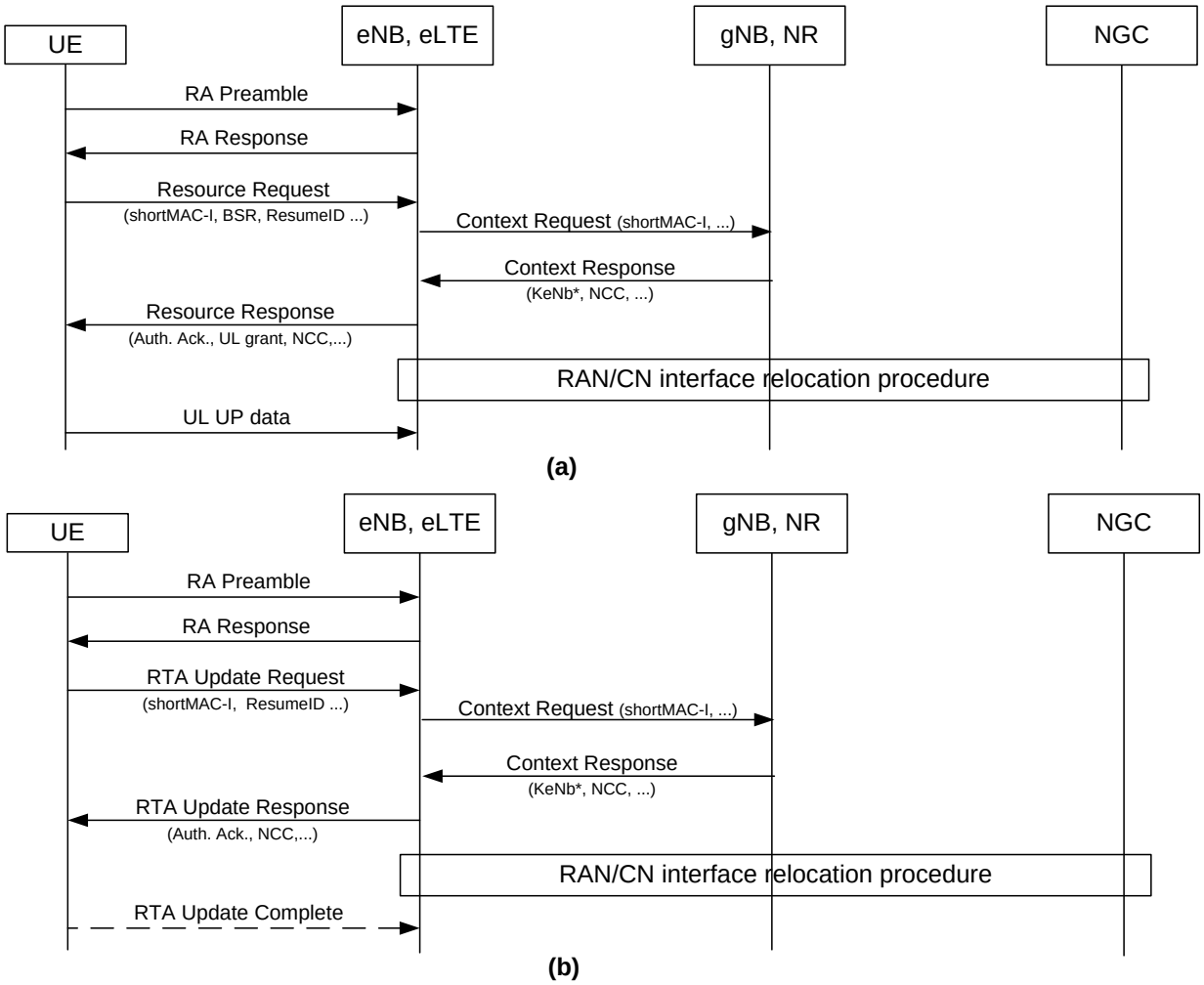


Figure 6-4: Inter RAT context fetching procedure as part of a) state transition and b) RTA update procedure.

Inter-RAT cell reselection with location update and context fetching

In this alternative approach, the UE is given a RAN notification area that consists cells that belong to only one RAT. Therefore, the UE need to initiate RAN notification area update procedure if it reselects a cell in a new RAT. However, rather than sending the UE to RRC Idle in the new state, it enters the inactive state in the new RAT with the UE's context is fetched from the anchor NB, as shown in Figure 6-4: b. In this approach, the UE does not need to autonomously change its RRC state during inter-RAT mobility, which would be relatively easier to implement. However, the RAN notification area update procedure consumes UE battery and incurs signaling overhead to the network.

6.3.4 Inter-RAT mobility during active state mobility

The control plane design for tight NR/eLTE interworking plays a major role in overall eLTE/NR interworking architecture design. Specifically, how the RRC state machine works for UEs with multi-connectivity towards both NR and eLTE still needs investigation and agreements in 3GPP specifications.

In the proposed solution, a master and slave RRC approach is considered, where there are two RRC connections from the network point of view: M-RRC that might be located at the LTE Master eNB (MeNB) and an S-RRC that is located in the Secondary node (SNG-NB). This is illustrated in Figure 6-5. The MeNB acts as the control plane mobility anchor from the core network point of view. A secondary node provides NR resources to the UE and the configuration of the SNG-NB is passed to the UE, transparently, by the MeNB at least for the initial configuration. The architectural details of NR/eLTE interworking are described in the Section 2.3.

The master/slave RRC approach has several mobility related advantages over the legacy approach. There is no additional latency introduced for direct signaling exchange between the UE and the SNG-NB. NR specific measurements could be reported to NR RRC directly. Mobility within NR maybe controlled directly by NR RRC and failure handling the SNG-NB is handled by NR-RRC. This allows for flexible operation and support of eLTE/NR tight interworking. The isolation of NR RRC and LTE RRC allows independent evolution of both protocols. Coordination is needed between M-RRC and S-RRC in cases where there are resources allocated for split bearer type of operation, band combinations supported by the UE and measurements. This may increase the implementation complexity on the UE side.

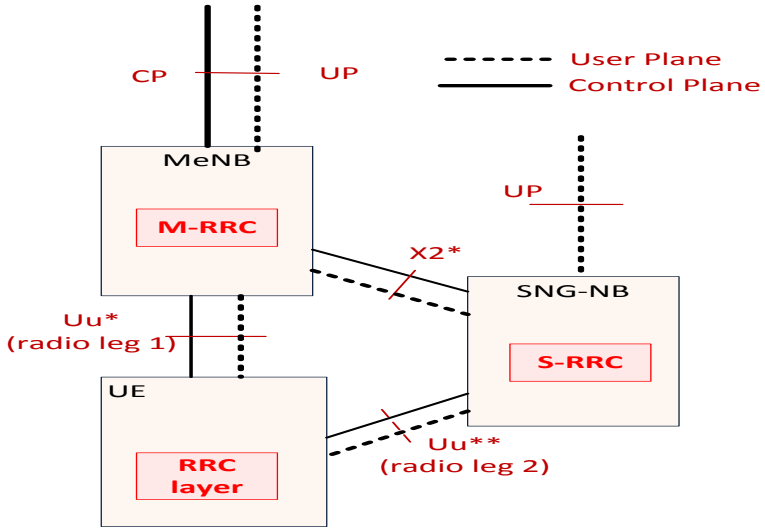


Figure 6-5 NR/eLTE tight interworking with NR and eLTE RRC

6.4 UE aspects of mobility

5G will efficiently support a diverse set of uses cases including eMBB, URLLC, and mMTC as well as functionalities such as access and backhaul and spectrum sharing, including coexistence with LTE. This creates many new design challenges related to usage scenarios, frequency ranges above 6 GHz, beam operation, numerologies and duplexing modes. Firstly, an overview on mobility from a UE perspective is given. Secondly UE capability signaling updates in LTE-5G tight-interworking scenarios were evaluated to investigate options for efficiently providing and aligning UE capabilities between LTE and 5G.

6.4.1 UE related mobility measurements and control

There are new use cases as eMBB, mMTC and URLLC with a diverging set of requirements. The extended frequency range above 6 GHz needs support of much larger UE channel bandwidth and carrier aggregation. Beam operation offers many options how the UE determines the best cell in particular taking antenna arrays and beams into account. Envisaged numerologies require support of flexible slot types and structures and duplexing modes in terms of flexibility must be taken into account. The design of radio resource management (RRM) measurements for 5G must avoid unnecessarily complex or restricted operation of different functions and services.

One of the key issues on 5G mobility design is whether to use the common RRM measurement for both IDLE and CONNECTED mode. Options are to use same RS resources (5G-SSS or 5G-SSS, MRS, 5G-SSS and DMRS for PBCH) or not same RS resources in IDLE and CONNECTED mode, which needs to be evaluated against cell coverage, overhead, accuracy and performance. In the following first investigations regarding 5G mobility procedures for inter-cell RRM measurements for CONNECTED and IDLE from a UE perspective required in a beamforming

system are disclosed. For example, the following transmission of synchronization signals forming a SS block, burst and set as shown in Figure 6-6 being made possible by antenna arrays and beamforming techniques is proposed supporting mobility measurements. From a UE perspective, SS burst set transmission is periodic. In such periodic transmission, the synchronization signals form a SS block, where multiple SS blocks get transmitted in series to form a SS burst. Multiple SS burst form a SS burst set, where the SS burst set repeats indefinitely.

Initial access of a UE exploits the support of its antenna array as well as the beamforming techniques. The number of beams the network cell wishes to cycle through depends on the deployment scenario and is a trade-off between the number of network beams and synchronization signal overhead. It is important too to limit the blind detections and measurement complexity from supporting too widely varying configurations of the synchronization signal. A good trade-off between network configuration flexibility and UE cell search complexity is to limit the burst periodicity and the SS burst set duration. The network configures a varying number of SS blocks within a SS burst such that the periodicity between any two SS blocks is always a fixed interval.

The number of SS blocks within a SS burst depends on the number of beams provided to transmit the SS burst. For cell deployments with many TRPs and many antenna array elements, the number of beam sweeps is getting very large within the SS burst set. Another drawback is the latency of RRM measurements in this setup since the UE must wait until all available beams in the network for each 5G cell are measured. This increases the cell re-selection time during IDLE mode.

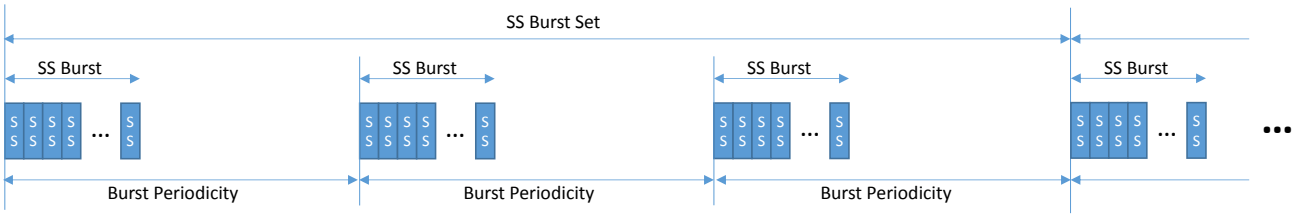


Figure 6-6 Periodically transmitted SS block, burst and set

Therefore, beam management and beamforming (Figure 6-7) for transmitting SS blocks needs to be done different compared to beamforming for data transmission. It is proposed to have a flexibility for setting up beams for SS blocks, bursts and sets adopted for IDLE and for CONNECTED mode. Then the network configures different RS for IDLE and CONNECTED mode UEs avoiding network overload by RS. If the same set of RS is needed for IDLE and CONNECTED mode UEs, all TRPs transmit SS bursts and the network can get overwhelmed. Differentiating these beams creates a huge burden on the UE and increase its power consumption during IDLE mode which shall be avoided as well. In the following a brief overview for DL and UL based mobility applying this setup is given.

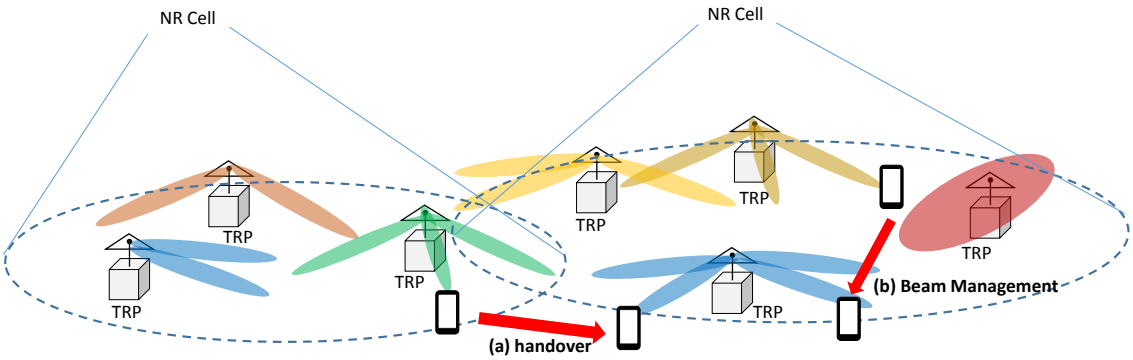


Figure 6-7 Mobility support for handover and beam management

In DL CONNECTED mode the UE is aware of RS from different TRPs and performs accordingly the measurement reporting. The UE detects the 5G-SS from different 5G cells performing time synchronization and measuring signals from different TRPs. Mobility in CONNECTED mode is supported by beam management. In DL IDLE mode, the UE surveys 5G-SS from surrounding 5G cells being ready for cell re-selection. Due to missing AS context and RRC connection to the network the UE monitors 5G-SS such as 5G-PSS, 5G-SSS, and 5G-PBCH. If the network configures a single 5G cell to a single TRP, it will be able to control cell re-selection during IDLE mode at TRP level. UL CONNECTED and IDLE mode mobility support to facilitate for example a fast handover or cell re-selection requires additional control and signaling but shall not sacrifice UE power consumption. Therefore, further investigations shall look firstly into 5G DL mobility developments avoiding duplication of functionality.

It is proposed to evolve CSI with a dedicated designed CSI-RS. Then one optional set of RS are 5G-SSS in IDLE and MRS in CONNECTED; 5G-PSS and/or 5G-SSS in IDLE; 5G-PSS and/or 5G-SSS and CSI-RS in CONNECTED. For example, in Figure 6-8 is shown, that a CSI-RS density of 1 RE per PRB with 1 OFDM symbol delivers similar RSRP data accuracy compared with SSS in low geometry and a much better RSRP accuracy in high geometry scenarios. Simulation setup was with 4 GHz carrier frequency and with CDL-C channel model with 500 ns RMS delay spread.

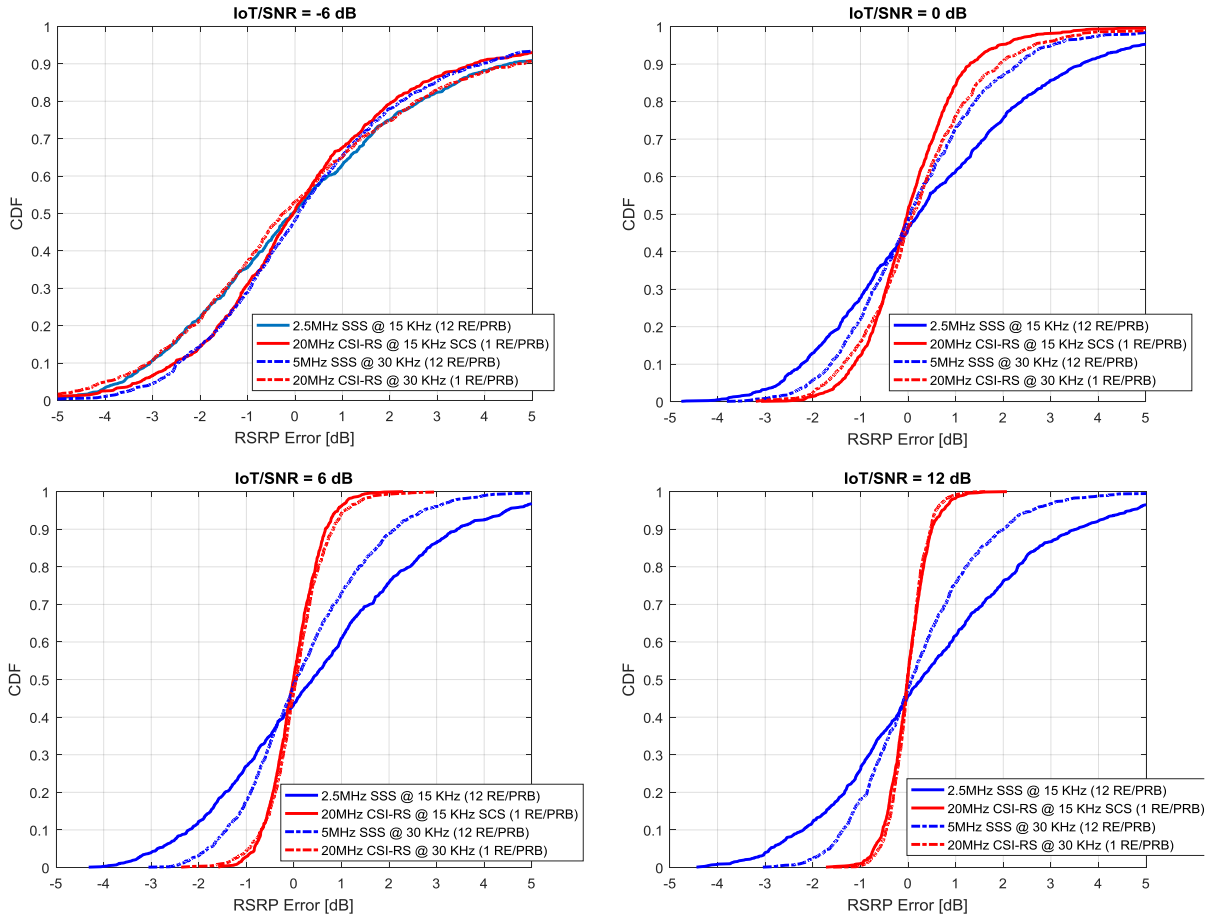


Figure 6-8 RSRP accuracy evaluation comparing use of CSI-RS and SSS

It is concluded that the RSRP accuracy measured with SSS may be sufficient in some scenarios and CSI-RS is only required in certain scenarios. Therefore, it is beneficial to let the UE specify the configuration of CSI-RS for L3 mobility and do this configuration separately from CSI-RS used for CQI/PMI/RI measurements.

6.4.2 UE capability signaling

In current 3GPP LTE-5G tight interworking scenarios, both LTE and 5G systems must provide configuration information to the UE of their RRC and protocol stack. However, some UE capabilities could be shared between LTE and 5G systems, and these capabilities were part of the research presented herein. Such shared UE capabilities still require a coordination mechanism between LTE and 5G NBs in order to assure that combined configuration of LTE and 5G stacks does not exceed the total UE capabilities. Therefore the different options for the coordination of UE capabilities, assuming tight interworking between 5G and legacy systems (e.g. LTE) were studied, where LTE-5G tight interworking scenarios (Figure 6-9 left non-standalone, LTE assisted EPC connect, right non-standalone, 5G assisted 5G-CN connected) were mainly investigated from the perspective of managing UE capabilities in different scenarios, e.g., D2D,

radio link failure and device state change. The possible implications on RRC design were investigated.

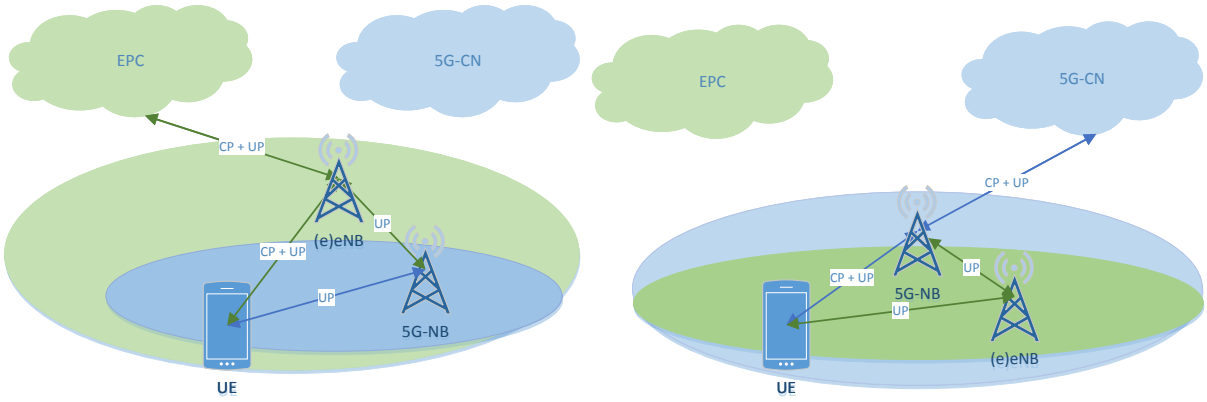


Figure 6-9 LTE – 5G tight interworking scenarios

The LTE-5G coordination mechanism shall:

- Enable an efficient distribution of shared UE capabilities between the LTE and the 5G stacks, e.g., with respect to the currently active use case, current network state, and current device state;
- Avoid adding a significant coordination signaling overhead, which may negatively impact user experience especially in fast mobility use cases;
- Enable an independent evolution of LTE and 5G stacks by avoiding the case where LTE eNB and 5G-NB are required to comprehend each other's UE configuration, and
- Aim to minimize differences between the UE capability configuration in LTE-5G tight interworking case and the LTE/5G standalone case.

LTE and 5G use LTE RRC and 5G RRC, respectively, to perform the UE capability coordination. Examples of UE capabilities, which need to be coordinated between eNB and 5G-NB are: i) PHY and RF parameters relevant to LTE/5G tight interworking needs, ii) the dynamic sharing of HARQ soft buffers between LTE and 5G or a semi-static split between LTE and 5G of the total number of soft-channel bits, iii) the 5G operation above 6 GHz as a single wideband carrier or by means of carrier aggregation as well as iv) a 5G UE flexibility like the LTE UE capabilities supporting UE implementation flexibility. Table 6-2 below shows more details.

Table 6-2: UE capabilities coordination

Category	Capability	LTE-5G interworking
RF	supportedBandCombination	Due to the potential RF sharing between LTE and 5G, some coordination is needed
PHY	Maximum number of DL-SCH/UL-SCH transport block bits received within a TTI	Considering the potential differences between LTE and 5G (e.g., TTI, channel coding), some capabilities will need to be shared between LTE and 5G

	Total number of DL-SCH soft channel bits	This is related to how LTE and 5G are implemented in UE, i.e., whether receiver buffer is shared
	p-MeNB and p-SeNB	This is related to the power allocation in UL. Some form of sharing of the total available Tx power between LTE and 5G is expected.
	supportedMIMO-CapabilityDL(UL)-r10	This is related to the maximum number of spatial multiplexing layers in DL/UL. Considering the potential differences between LTE and 5G (e.g., MIMO schemes and TTI), it is expected that the capabilities will not be shared between LTE and 5G.
PDCP	maxNumberROHC-ContextSessions	This is related to the maximum number of header compression context sessions supported by the UE. This depends on UE implementation and coordination between LTE and 5G is expected.

To achieve this in a LTE-5G tightly integrated network scenario, a first solution is, that LTE and 5G-NBs and UEs implement and comprehend both RRCs, and the capability container is fully comprehended and verified before the configuration is sent to the UE. This solution is not only costly, but also makes inter-vendor inter-operability between LTE and 5G quite complex; and an independent evolution of LTE and 5G quite difficult, increasing significantly the complexity of NBs as well as UEs.

The solution proposed is to include an additional RRC container between eNB, 5G-NB and UE for fields that need coordination. This is additional to the container carrying already the RRC configuration provided to the UE. This allows 5G-NBs to provide 5G configuration directly to the UE over 5G without having to send it over LTE eNB. This setup does not require coordination and is sent over 5G radio to the UE without any involvement of LTE eNB, which makes it much faster. Only when there is a change of parameters that needs coordination, the LTE eNB will get involved. It is sufficient to send just the container with coordination parameters to LTE and 5G can still configure the UE directly over 5G.

Furthermore, it is possible to do the coordination upfront. That is, if 5G wants to do some configuration, it can do the coordination in advance. When it actually does the UE configuration, it can do so directly without involving LTE eNB. The benefit of sending two containers with complementary information is that it is easier to implement in the UE. Internal to the UE, it will ease implementations as it knows which parameters need coordination and only needs to handle the coordination of a small set of parameters rather than the full configuration.

There are several implementation options. Firstly, the information in this additional container is a “duplicate” of the information in the full configuration container. Secondly, the full UE configuration is split between the two – that is, both containers are sent to the UE and then the UE puts together the full configuration from data in both containers. Thirdly, the receiving NB may combine the information from the two containers and send a single container to the UE.

LTE and 5G-NBs must still check the information in the container to ensure compliance of the configuration to the UE capability. If the configuration is not valid or compliant with the UE

capability, the receiving node can reject the configuration request and it will not be sent to the UE. LTE and 5G-NBs must be able to understand all information in the container for the configuration to be successfully accepted. If any of the two systems does not understand any of the fields, it indicates that the network implementations are out of sync with respect to what needs to be coordinated, and can result in a wrong configuration to the UE. The split of UE capability data between the nodes can be done ahead of the configuration between the two nodes or even after the configuration. This split can be done by either node and sent to the other node also using containers. The capability related containers could be part of the same container as configuration or different from the configuration container.

The data in the proposed additional container could be logically categorized into the two groups. Firstly, a shared resources group (SRG), examples of the parameters in this group are the supported common bands and the maximum number of robust header compression (ROHC) context sessions. Secondly an activity collaboration group (ACG), examples of the parameters in this group are the supported band combinations and the maximum power allocation in UL. In order to design an efficient LTE-5G coordination mechanism, the management of the UE LTE/5G shared resources could be done in the two steps, firstly the initial split of the UE LTE/5G shared resources and secondly a dynamic update of the UE LTE/5G shared resources split.

It is sufficient to provide configuration from one node to another for example just from 5G-NB to LTE eNB rather than mutual exchange of configurations between them. Without providing a node with the other's relevant configuration, it would not be possible for the node to ensure that its configuration does not violate the other's configuration. For example, 5G-NB needs to know the eNB configuration to ensure that its configuration is compatible with UE, and vice versa. Hence the container must be defined in LTE RRC for LTE configuration, and 5G RRC for 5G configuration to be exchanged between the network nodes.

When the 5G NB wants to configure the UE, it generates two containers. One with parameters that need coordination and another with parameters to be sent to UE. It then sends both to LTE eNB. The LTE eNB then checks only the coordination container to verify if the configuration is compatible with UE. If so, it forwards the other container with UE configuration to the UE. The same can be applied for 5G-NB. This is shown in Figure 6-10.

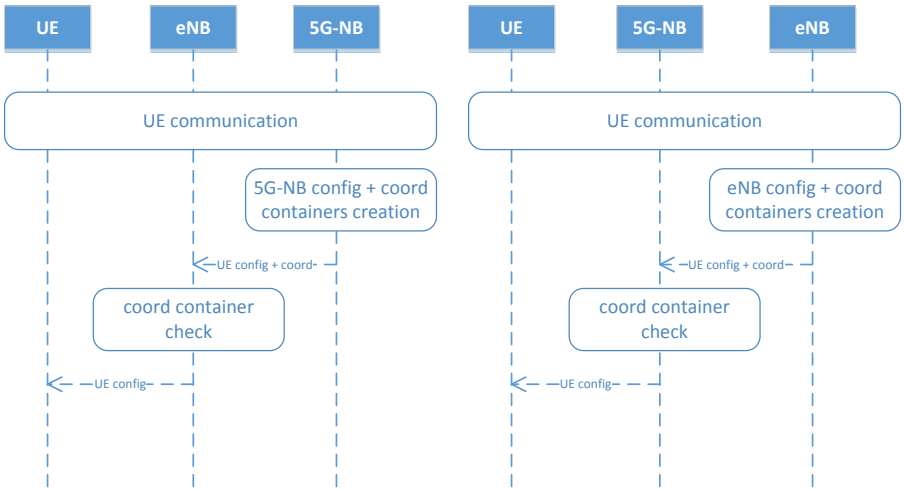


Figure 6-10 eNB – 5G-NB tight interworking container handling.

The container split proposed can also be used for cell change and simplified for no configuration changes. If there are no configurations that need to be verified by LTE eNB, 5G NB only sends one container that is to be forwarded to the UE and no specific action is needed by LTE eNB. It is also possible to send the UE configuration directly from 5G-NB to UE over 5G radio. When LTE eNB needs to reconfigure the UE with parameters that need coordination, it provides a container with those parameters to 5G NB.

As part of the coordination container checks, LTE eNB and 5G-NB ensure that they can comprehend all of the fields included in the container. If at least one of the fields cannot be comprehended, the container data shall get rejected for becoming part of configuration and it indicates there is a mismatch in the LTE and 5G capabilities for the parameters that need further coordination.

One further solution to avoid that a LTE eNB must implement the 5G RRC is to use LTE RRC to provide an “equivalent” configuration as the 5G configuration enabling it to check for UE capability violation. It assumes that there is the capability to mapping a 5G configuration onto a LTE RRC configuration. Further study will be needed to evaluate such a solution.

6.5 Context aware mobility

Mobility prediction plays an important role in designing of context aware radio resource management (RRM), which aims at providing uniform service quality. Knowledge of future user location (position, route or next cell) can be used to anticipate future data traffic conditions, future events (crowd formation, traffic jams etc.) [KKS+13] [KKS15] and appropriately reserve or manage resources to provide optimum service. In real world scenarios, user mobility is not random but is rather direction oriented. The user direction relies on its origin and destination. Further, there are several users who exhibit similar mobility patterns on daily basis (e.g., office goers, public transport). They tend to regularly traverse a limited set of trajectories, comprising of specific landmarks. Such mobility can be referred to as Diurnal mobility, which constitutes a major

portion of mobile users. In this work, users following diurnal mobility are considered and information arising from such mobility (e.g., origin, landmarks, destination) are used to enhance the accuracy of mobility prediction. In day to day life, there are several instances when a user will run into a coverage hole (e.g., tunnel), where his throughput will be nil. By anticipating the encounter of such a coverage hole in near future, it is possible to allocate more resources suitably and buffer the data before running into coverage hole. In coverage hole, buffered data can be used to sustain streaming/full buffer service for non-real time services. This will enable a uniform service experience for the user, even in deep shadow regions or coverage holes. Elaborate details about the functioning of this technology component can be found in [MII16-D61].

This TeC makes use of [MII16-D21] as evaluation methodology basis. Madrid Grid test environment is used as the evaluation scenario. There are coverage holes present in the simulation scenario at different roads as shown in Section A.9, where the achievable throughput is zero. There are 12 micro BSs with LTE-A technology (bandwidth of 10 MHz, 50 physical resource blocks (PRB) at 2 GHz carrier frequency). User is allocated with 1 PRB. The context aware resource allocation is triggered by route prediction, by allocating more resource blocks (PRB = 2 or PRB = 3) and buffering data. The simulations are carried out for 100 runs. A user is assumed to follow diurnal mobility and would traverse among 10 known trajectories with different probabilities. Context information about origin, destination of users, roadmaps with coverage holes is assumed to be known at a co-ordinate level. More explanation on considered set up and mobility model could be found in [KZS16]. With the proposed approach the accuracy of next serving cell and next route prediction in the considered scenario is increased to 85% and 90%, respectively, as opposed to 40% and 60% accuracy with simple Markov model. Comparison with next cell prediction schemes based on user geometry (dB) and distance metric [KKS+13] [KKS15] are also given.

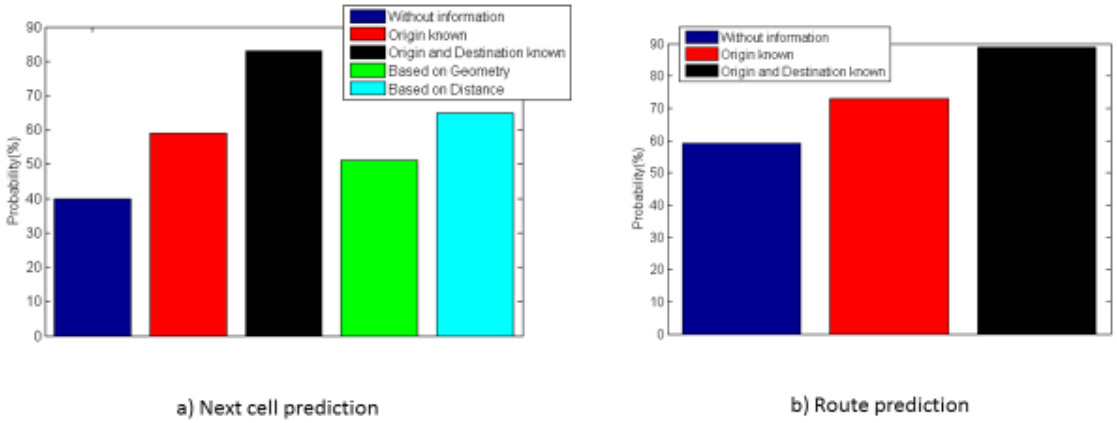


Figure 6-11 Mobility Prediction Results

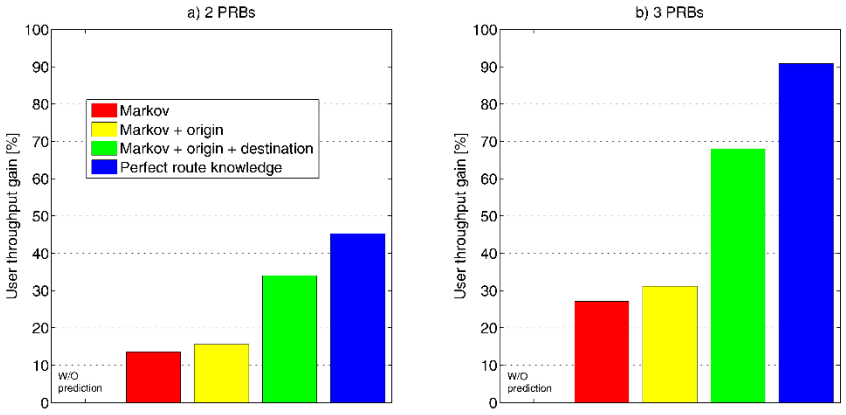


Figure 6-12. Throughput improvement in routes with coverage hole.

From Figure 6-12 it can be seen that context aware resource allocation triggered by enhanced route prediction improves the throughput in the routes with coverage holes. The UE is allocated originally with one PRB and as soon as a coverage hole is anticipated via route prediction, additional PRBs are allocated.

6.6 Data Analytics based Traffic Engineering

As thoroughly described in [MII16-D61] the use of context information is imperative for enabling the efficient use of the network resources [MSC13]. However, the use of context comes with a certain cost in terms of signaling and delay for making the proper decisions. In particular, several Context Aware mechanisms are available in the literature ([XPM+14], [CWH+13], [KKP+13], [QXY+10] [MBS+10], [MCC+11].) which however mainly rely on the use of real time information which has to be collected and processed. Additionally, even enhanced with state optimization, the rule based systems (typical approach for context enhanced schemes) cannot address the "Curse of Dimensionality" ([SRG+12]) efficiently and in real time [MKS+10].

One additional aspect of the context aware mechanisms is the incorporation of the user preferences. Regarding the latter, it is not clear how they are introduced to the system. Several literature proposals suggest that the users should introduce their preferences manually. In 3GPP, the user behavioral profile may be used for deriving UE specific cell reselection priorities to control idle mode camping or for deciding on redirecting active mode UEs to different frequency layers or RATs [3GPP15-23060], but it is not specified how this behavioral profile is built. Additionally, it is static for each user and captures his generic behavioral preferences and it is not linked to special contextual information (e.g., location, mobility, user device, etc.). Thus it would be really beneficial if the operator could identify the user preferences automatically, using UE statistics, and combining the profile with additional information, such as the UE location, mobility, battery level, etc.

Furthermore, the users tend to change their behaviors depending on the location, the time, and other characteristics as well such as user equipment type, the battery level of the user equipment,

the charging status of the user account (e.g., remaining credits, etc.), the overall user income, the user educational education level, etc. Due to the complexity of the future networks, it is imperative to identify such correlations and take advantage of them to better manage the network resources.

A simple example of modeling the user behavior depending on the user context could be the following:

- Joe at the Office every weekday from 9:00 – 18:00, is stationary, he performs long voice calls, and he does not access internet through his cell phone.
- Joe at his House every weekday after work is stationary, accesses web applications via his cell phone using WiFi, and he does not make phone calls.
- Joe in city center every Saturday from 10:00 – 16:00 is highly mobile, he performs short voice calls, and he does NOT access the Internet through his cell phone.

Such knowledge could enable the network to predict the overall throughput requirements in certain locations for specific time periods and proceed in the relevant actions in advance. For example, one network operator may rent certain spectrum for a certain period of time, if certain users are located in a mall.

By analyzing the past user actions (past behavior) with data mining mechanisms, we may identify the most impacting parameters in the user behavior. Afterwards learning mechanisms can build user behavioral profiles. These profiles could be used for predicting the overall throughput requirements in certain areas, or even be used for optimized mobility management, call admission control, etc. For the aforementioned steps, state of the art mechanisms may be used. In particular, for the identification of the most impacting parameters information gain (IG) can be used to measure of how much information the presence or absence of a particular term contributes for making the correct classification decision, using entropy measurements. Further options to be used for feature selection are the Frequency-based feature selection, Chi square, etc. For the learning part these solutions supervised or unsupervised learning may be used; examples of supervised learning schemes used for this can be Decision Trees or kNN whereas for unsupervised learning schemes we may use typical clustering schemes (e.g., K-Means).

The dimensions of the profile could be rough location(e.g., provided during the TAU/initial attachment)/day/time as well as battery level and charging status. This approach has two phases of operation, the offline one and the online one. In the offline phase the required inputs are being gathered and processed, so as to extract the behavioral profiles. Knowing the user behavior with high accuracy will further facilitate the network operator to proceed in the respective actions such as small cells placement, spectrum acquisition, etc. Furthermore, this will enable the operators to optimally utilize the available spectrum (e.g., by offering it to other operators with certain pricing) or even proceed in spectrum acquisitions which will be more suitable for the users in a certain location with specific service requirements (e.g., acquire spectrum using different spectrum authorization options). Finally, the user profiles can facilitate the operator to predict the user requirements for SON approaches (energy saving, data pre buffering).

Using the offline extracted profiles will enable the reduction of the signaling exchange. In particular, the UEs will inform the network under for:

1) Deviations from the predicted behavior
2) Changes in real time information – e.g., battery level

Figure 6-13 shows the signaling cost for updating the RAN for 25 UE in a considered area providing updates regarding their position, accessed service, mobility, battery levels for a certain time window (in this case 180 minutes). Three schemes are considered: (a) a context aware scheme where the RAN is informed periodically every minute, (b) a context aware scheme where the RAN is informed periodically every minute but with less accurate location information, (c) the profile based scheme where we have various compliances to the profile. Every time that a UE deviates from the profile, the network should be informed to act accordingly; thus, in one case we consider that the UEs comply with their profiles 50% of the time regarding mobility, in the second one we consider that the UEs comply with their profiles 50% of the time regarding their mobility, and in the third case we consider that the UEs comply with their profiles 90% of the time. As it is shown the gains are significant. If we consider even higher concentrations of UEs even more significant gains can be achieved, highlighting the gains in the RAN.

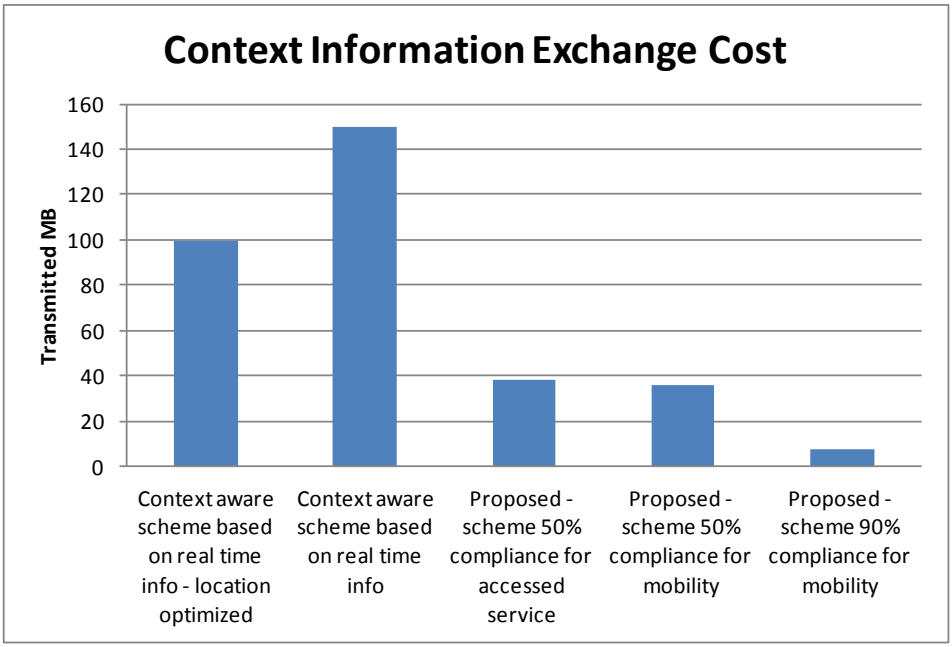
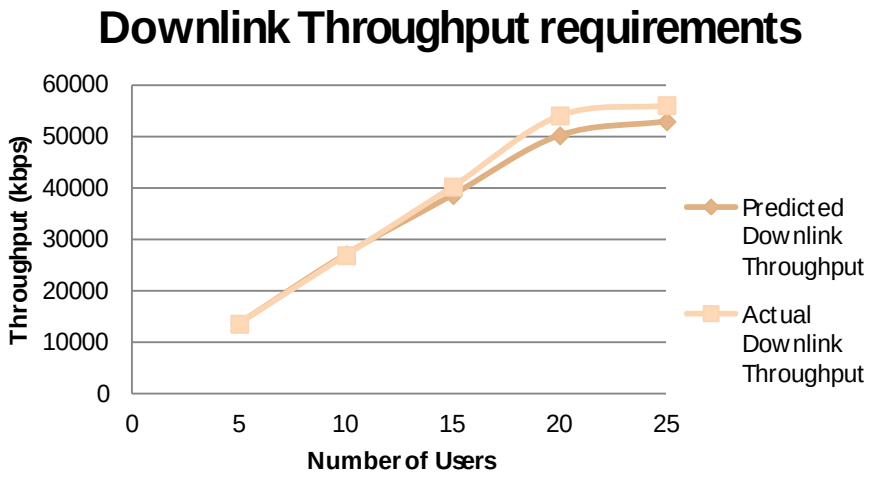


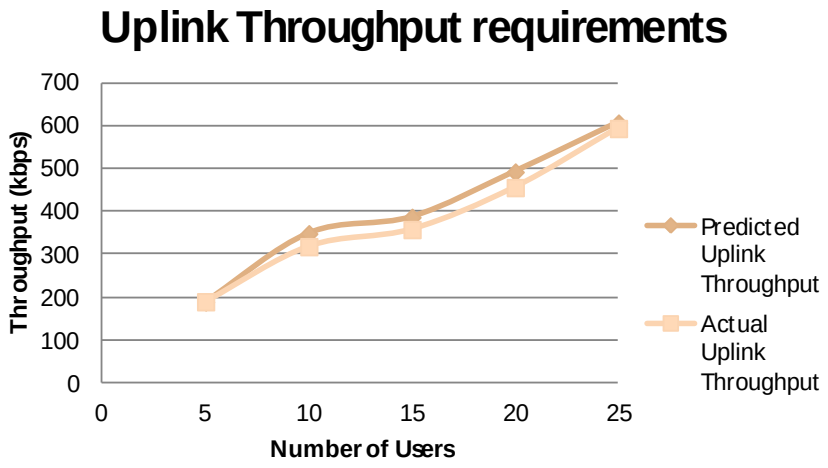
Figure 6-13 Context Information Exchange Cost

The use of profiles for predicting the based on previous behaviors of course facilitates the accurate decision. However, it is not ensured that the UEs will follow their profile all the times. In these cases the network will operate using the typical mechanisms for admission control and handover. In order to ensure that the prediction mechanism is accurate enough, we have performed a set of simulations where the 10% of the users randomly changes its behavior (accessed service or mobility, or both). The simulation setup is a single floor Mall environment (2

macro eNBs, 3 home eNBs, and 1 GSM macro cell) with users who access VoIP, Video streaming, Web, and FTP services. Even in these cases the prediction mechanism can identify the uplink and downlink resources properly as shown in the Figure 6-14. In total the traffic demands can be predicted very accurately; in particular we observe less than 5.2 % average error and less than 4% standard deviation.



(a)



(b)

Figure 6-14 User downlink (a) and uplink (b) throughput requirements prediction using profiles

The process of storing and distributing the UE profiles up to now has been performed using the 'Index to RAT/Frequency Selection Priority' (RFSP index) which is used to describe the user preferences in terms of accessed services and mobility [3GPP15-23060]. The extension of the

RFSP so as to capture the various UEs behaviors as well as the requirements prediction has a direct impact in the RAN. The RFSP index is able to describe only a 256 behaviors and lacks flexibility thus making the description of the user behavior very static. In particular, the RAN should be updated in such way so as to describe with higher granularity the user predicted behavior. With the multi-dimensional profiling the user behavior will be accurate and the network actions more targeted.

7 D2D

7.1 Introduction

Device to Device or D2D communication refers to “direct mode” or “locally routed” path for communication between UEs. Direct mode here refers to scenarios where devices participating in D2D communication are at the same level of network hierarchy and are located in proximity to each other. This chapter provides details of several of the D2D considerations and challenges from CP design perspective e.g. to support efficient and optimal resource management and channel access, cooperative communication, cellular coverage extension, network offloading and inter-RAT/intra-RAT Mobility. This chapter is, to a good extent, a continuation and further in-depth work of D2D control plane specific work discussed in METIS II Deliverable D6.1 [MII16-D61]. To be more specific, for example, cooperative D2D communication as proposed here, along with proposed interference management, significantly improves spectrum efficiency. Proposed mobility management scheme satisfies D2D service continuity requirements ensuring that UEs in the D2D/V2V group (of two or more) are handed over to the target RAN without breaking the D2D/V2V link while also enabling URLLC. Besides, taking into account different context information, D2D communication can reuse the resource of cellular users for its transmission and therefore solution proposed here significantly improves the network capacity.

7.1.1 State-of-the-art

D2D concept has been continuously evolving both from academia as well as industry perspective. Various flavors of D2D has been there since quite some time e.g. WiFi Direct, Bluetooth, etc. Also, 3GPP started developing D2D specifications, also called as Proximity Based Communication or ProSe, as part of its Release 12 and Release 13 work items. 3GPP work focused on two aspects: ProSe discovery and ProSe communication and they are specified in [3GPP16-23303]. Most of the work so far has been focused to support public safety scenarios. For public-safety uses, 3GPP has specified protocols to enable UE-to-Network Relays, thus enabling out-of-coverage UE to communicate via UE-to-Network Relays. [3GPP15-36211] specifies RAN aspects including synchronization signal design and synchronization procedures, type 1 and type 2b discovery, physical layer design for discovery which includes resource allocation and discovery signal design, mode 1 and mode 2 communication, L3-based UE-to-Network Relay and D2D for inter-frequency and inter-PLMN discovery. In case of type 1 discovery approach: resources are allocated on a non-UE specific basis, applicable to UEs in connected or in idle mode, Tx resource pool is provided in S-SIB or RRCConnectionReconfiguration message and Rx pool is provided in S-SIB. In case of type 2b discovery: resources are allocated on a per UE specific basis, applicable to UEs in connected mode, Tx resources are provided in RRCConnectionReconfiguration message and Rx resource pool is provided in SIB. In [3GPP17-36331], RRC protocol has been extended to cover certain control plane design aspects for sidelink, i.e. sidelink discovery and communication monitoring, sidelink discovery and communication transmission, sidelink synchronization process. Currently, 3GPP has just started,

as part of its ongoing Release-14 scope, work on evolution of ProSe including new aspects such as V2X and D2D in wearable and MTC devices.

7.1.2 D2D framework for 5G

In order to support several new possibilities envisioned in 5G, D2D is expected to play an important role as an integral component of 5G system. Going forward into the 5G era, D2D functionalities are expected to be natively supported into the CP protocol stacks of novel AI or AIV, rather than as add-on functions, from device as well as network perspective. This requires addressing several challenges from a CP design perspective. Important D2D aspects considered here for a 5G framework are the following:

- Channel sounding and control signaling among pairs of devices
- Cooperative D2D communication
- D2D discovery and communication to assist uplink transmission of remote UEs
- Self-backhauling
- Cellular coverage extension using D2D relays for mMTC
- Network offloading
- D2D inter-RAT/intra-RAT Mobility Management
- Efficient sidelink resource management and channel access
- Context aware group mobility management
- D2D enabled group based RACH access

METIS II has been working on and developing following solutions, some of these are discussed in METIS II deliverables D6.1 [MII16-D61] and D5.2 [MII17-D52]. Below sub-sections provide details of these and other related aspects of D2D in 5G:

Channel sounding among pairs of devices. One key requirement in the context of D2D communications is the need to estimate links between devices. The METIS-II assumption is that this should be done based on reuse of the same sounding reference signals (SRS) that are also used for the cellular uplink. This is different to the LTE Rel. 12 approach, where dedicated signals are used e.g. for link estimation. A key challenge is then that a device can either send its own SRS or receive SRS transmitted from another device but not do both at the same time. Hence, it is required to design SRS muting patterns such that devices can estimate the links to other devices in proximity over time.

Control signaling among devices. Another difficulty is how to enable control signaling between directly communicating devices (e.g. ACK / NACK, channel quality indicator, CQI, feedback etc.); in particular, if this is expected to build upon the same control channels as designed for cellular communications. If for instance there are certain signals foreseen for uplink control signaling, and others for downlink control signaling, then if the uplink control signals are reused for the control signaling between D2D pairs, a device can only transmit uplink control signals or receive these from another device, but not both at the same time. Solutions are here to either again apply a

muting pattern to the control signals, as in the case of channel sounding, or to relay control signals via an infrastructure node.

Cooperative D2D communication is when D2D pairs are utilized as relays to facilitate the transmission between a cellular user (CU) and its base-station (BS) to improve spectrum efficiency. In this case, the PC5* interface is enhanced to support unicast D2D communication and / or one-to-many / one-to-all D2D communication among pairs of devices, where one of these devices can be the source (DT or D2D Transmitter), while other devices can be the destination (DR or D2D Receiver). Besides, such D2D devices facilitate cellular user transmission by acting as relay devices. A cooperative communication scheme as proposed here enables 5G RAN to dynamically allow cooperative D2D mode selection and communication, while at the same time ensuring interference mitigation e.g. in case of simultaneous D2D and CU to BS communication over the shared radio resources, etc. More precisely, three types of cooperation are considered, namely overlay, underlay and hybrid cooperation [MII16-D61].

D2D discovery and communication to assist uplink transmission of remote UEs. In a D2D discovery procedure, network performs the D2D pairing algorithm by analyzing dynamically collected context information. After receiving network configuration, a discovery announcement message is sent together with a reference signal to the discoveree i.e. UE which is being discovered. The discoveree UE then responds with an ACK/NACK based on its calculated RSRP of the D2D link and other consideration. In a D2D communication procedure, a remote UE transmits its data packet to the corresponding relay UE with which it has been paired in the D2D discovery procedure. A random access procedure and D2D link configuration procedure might occur in this step based on whether the uplink traffic from the remote UE is periodic or not. Afterwards, the relay UE will forward the successfully received packets to the serving BS. The proposed D2D discovery and D2D communication procedures are also supported in connected inactive UE state. It is only necessary for a relay UE to enter RRC active state if the relay UE needs to forward the result of a discovery procedure or to forward a data packet of a remote UE to the BS. The required configuration information to support D2D operation is carried by sidelink system information blocks (S-SIBs) and downlink control information from physical downlink control channel (PDCCH). Moreover, when paired D2D UEs stay in connected inactive state, certain context information related to the D2D link can be kept in both relay UE and remote UE(s), in order to reduce both signaling load and power consumption for D2D communication.

Self-backhauling. A basic functional requirement for self-backhauling is the need to be able to align the transmissions on backhaul and access links from the perspective of a self-backhauled entity. If, for instance, a self-backhauled node uses one transceiver for both access and backhaul, it must be possible to multiplex the control signaling on the backhaul and access links in time, meaning that both the backhaul and access link must be able to be configured to use certain muting patterns w.r.t. control signaling. If a self-backhauled node has separate transceivers available for backhaul and access, this half-duplex constraint would be relaxed. However, it may still be necessary that the node synchronizes the usage of transmission and reception on the backhaul and access links (i.e. it may not transmit control signals on the backhaul link while it is receiving control signals on the access link, due to potentially too large cross-interference).

7.2 Context-aware D2D communication to serve mMTC

mMTC devices can be located everywhere, even in the deep indoor placements. To overcome the propagation constraints in mMTC communication and related power dissipation challenge at device side, METIS-II study the exploitation of context-aware D2D communication for mMTC [MII16-D61] that could provide improvements, in terms of service availability and battery life of mMTC devices. In this particular work, certain UEs are selected by the network to act as relay UEs for mMTC devices located in cell boarder or in deep indoor. In order to optimize the system performance in terms of service availability and device power consumption, context information are collected and exploited by the network to efficiently set up D2D pairs.

7.2.1 System model

Figure 7-1 provides a graphical description of this scenario. As it can be seen from this figure, sensor #4 and sensor #5 experience bad channel conditions for their cellular links and thus are referred as remote UEs. Meanwhile, sensor #2 is seen by BS as an optimal relay node for sensor #4 and #5. Thus, D2D connections are established between sensor #2 and sensor #4, and also between sensor #2 and sensor #5. After that, uplink data of sensors #4 and #5 are transmitted to BS through sensor #2. Besides being a relay node for remote UEs, sensor #2 also transmits its own packet to BS. Moreover, sensors #1 and #3 are configured as normal cellular UEs and they are only responsible for transmissions of their own packets.

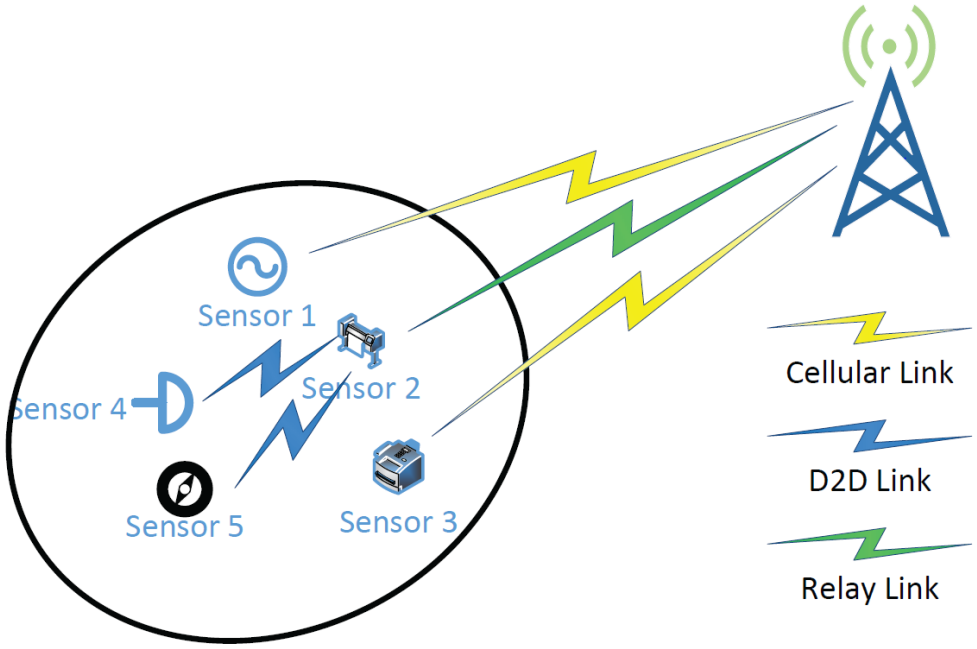


Figure 7-1 Exploitation of D2D communication for mMTC

7.2.2 Context-aware virtual D2D clustering and transmission mode selection

As mentioned before, D2D communication is exploited in this work to facilitate the uplink reports from remote UEs. In this scheme, three different transmission modes exist, as following:

(1) cellular transmission mode, in which the devices upload their reports to BS with cellular links;
(2) relay transmission mode, in which the devices are configured by network to relay the reports from remote UEs and meanwhile transmit their own reports to BS;
(3) D2D transmission mode, in which the remote UEs transmit their reports to relay UEs.

In order to adapt to system changes in real time, MTC UEs are dynamically configured with their transmission modes by the BS. Thus, the proposed context-aware D2D scheme can be divided into two steps:

1. clustering sensor devices into different groups;
2. selection of transmission mode for each UE.

Virtual D2D clustering

In order to achieve a high efficiency, D2D communication should only be used between devices that are located in the proximity of each other. Thus, BS should appoint a relay device from the neighborhoods of the remote device. An efficient approach to implement this scheme is to perform virtual D2D clustering, whose target is to group different devices together so that only the devices inside the group can communicate with each other through D2D communication. Four different methods are introduced in Annex A.4.2 to group sensor devices with different considerations. Moreover, for sensors locating near the BS, the signal propagation losses are relatively low and thus the exploitation of D2D communication for these UEs is not considered.

Transmission mode selection

In order to optimize the system performance, a proper selection of transmission mode for each UE is also critical. Thus, a smart transmission mode selection (TMS) algorithm should be implemented in the BS, taking into account of the collected context information. The context information includes all information related for TMS, e.g., channel state information (CSI) between BS and UEs, location and battery level information of UEs. The task of TMS is to configure devices in each cluster so that each device is aware of the transmission mode it should apply for its uplink report. Multiple context information is taken into account in this work to achieve an efficient TMS algorithm. The transmission mode selection of a device to set up D2D connections or to serve as a relay node is described in Annex A.4.3.

Once the BS obtains the list of feasible relay UEs in one cluster, BS picks up one relay UE and sends the D2D setup command to both the relay UE and remote UE(s). Upon receiving the D2D setup command, channel conditions between the relay and remote UEs are estimated to inspect if the D2D communication can contribute to a better energy efficiency. If a D2D setup procedure

is successful, the established D2D link is exploited for uplink transmission of packets from remote UE. The corresponding signaling schemes are detailed in Annex A.4.1.

7.2.3 Evaluation and Numerical Results

In order to evaluate the proposed technologies, a system level simulator is implemented in this work. Figure 7-2 shows the considered environment and deployment models in an urban area. Detailed information regarding the used models can be found in Annex A.4.4.

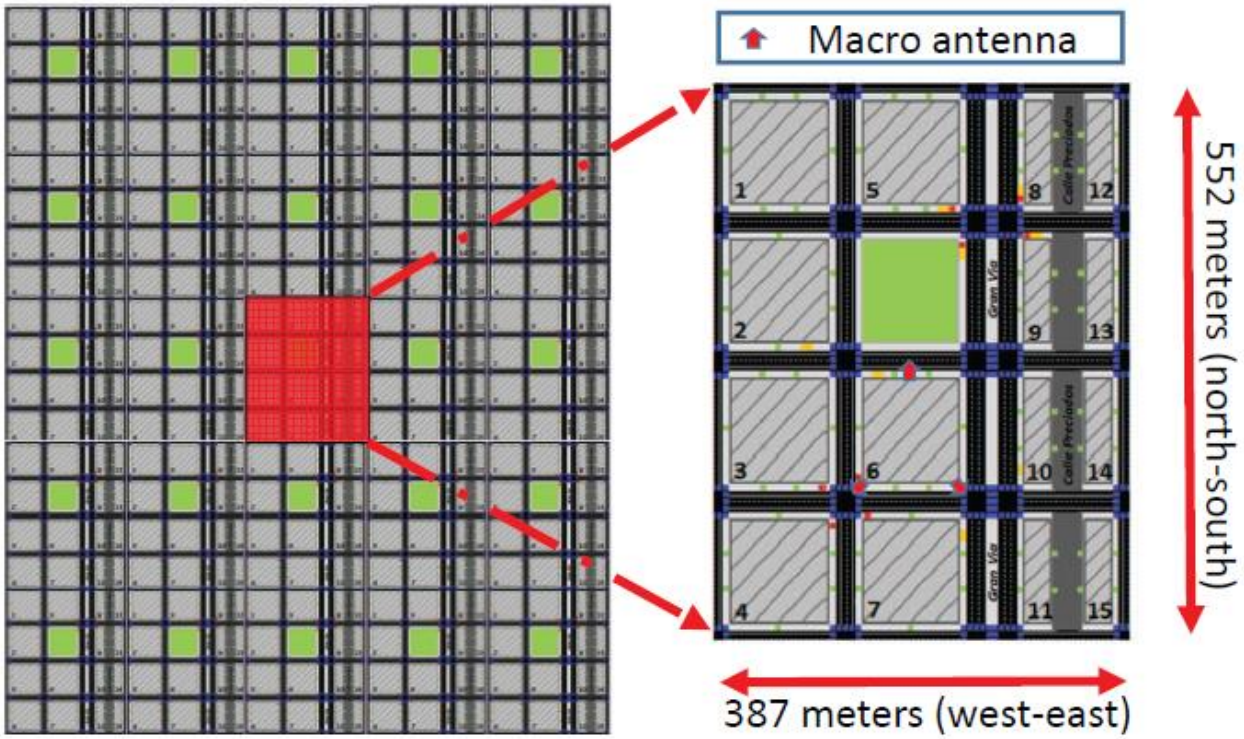


Figure 7-2 Environment and deployment model

In Figure 7-3: the cumulative distribution function (CDF) of served days for mMTC UEs is shown and the performance of LTE system is also drawn. As an input for geometrical clustering method, the area of a cluster is set to be 2500 square meters. Meanwhile, as an output from geometrical clustering method, the number of clusters is further fed to other three clustering schemes, in order to achieve a fair comparison among different schemes. In this work, LTE technology is used for modeling radio links. Thus, if a radio link experiences a very bad SNR value, no data transmission is possible on this link and the user is in network outage. As can be seen from Figure 7-3: , 15% of mMTC UEs are in outage of LTE network and cannot be served (w.r.t. the UEs with served days of zero). The steps in the figure is due to the fixed modulation and coding scheme (MCS). Due to the fixed MCS, though some UEs have different SNR values, their spectral efficiency can be same if their SNR difference is inside a small range. Therefore, the transmission time to carry one data packet for different UEs can be the same. And if the maximal transmission power is achieved, the power consumption for these UEs are same. By applying the “Distance+CSI based clustering” scheme, 99% of UEs can be served by either cellular or D2D link. In addition, 75%

UEs can meet the 10 years battery life requirement in LTE network, while this value can be improved to be 90% by our proposed scheme. Moreover, the performance of the users in outage of LTE network is given in Annex A.4.5. As shown by Figure A-12 regarding UEs who are in outage of LTE, 97% of them can be served by D2D communication and 55% of them can even be served for more than 10 years (3650 days) by their equipped batteries. The performance of UEs which are in coverage of cellular network is also provided in Figure A-13. In that figure, 87% of UEs can meet the battery life requirement of 10 years in LTE system, while this value has been improved to 96% by using our D2D scheme.

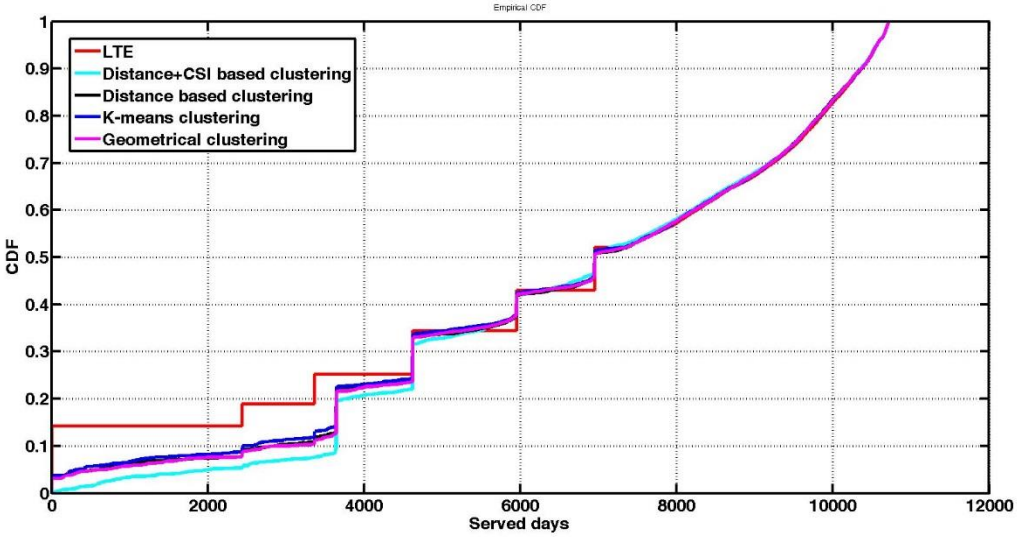


Figure 7-3: CDF plot for served days of mMTC UEs

7.3 Context-aware D2D underlay to improve system capacity

7.3.1 Introduction and system model

In the literature, cellular networks that utilize D2D communications offer several advantages compared with traditional cellular network without D2D, i.e. lower end-to-end latency and higher spectral efficiency [DRW+09]. In order to realize the proximity, reuse and hop gains, the available spectral resource need to be allocated for D2D communications in an efficient way. Presently, there are three modes of D2D operation that can be envisioned:

- Reuse Mode: D2D devices directly transmit their data by reusing some resources of the cellular network. The spectrum reuse can be either in uplink or downlink communications.
- Dedicated Mode: The cellular network dedicated a portion of resources for D2D devices for their direct communications.

- Cellular Mode: D2D traffic is relayed through eNB's in the traditional way.

In order to improve the efficiency of valuable spectrum resources, the reuse mode for D2D is investigated where the same spectrum resource is shared between D2D and cellular users. The reuse mode becomes important when available frequency range is considered as precious and it is essential to reuse the same spectrum resource. In this paper, it is assumed that D2D communications operate in reuse mode.

We consider a scenario where mobile users are divided into two categories, cellular users and D2D users. Cellular users request communication service directly from a BS. In comparison, two nearby D2D users form a D2D pair and perform a local communication service in order to exchange data with each other. Since in our scenario D2D communication in reuse mode is assumed, there are no dedicated resources available for D2D communication. Hence a D2D pair can only reuse resource block of a cellular user. Further, it is assumed that the uplink resource of cellular user can be reused by maximally one D2D link in one cell sector. The detailed system model and the corresponding RRM is provided in Annex A.5.1.

7.3.2 Signaling schemes to support context-aware D2D communication in reuse mode

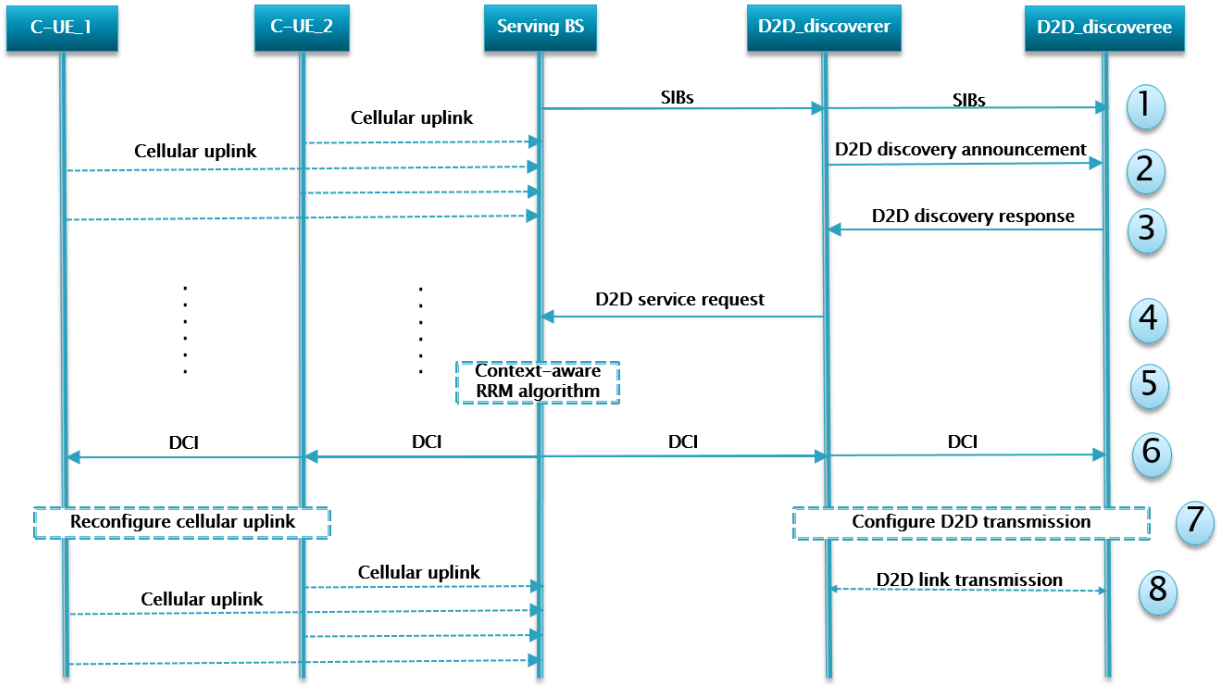


Figure 7-4 Signaling scheme in single cell to enable context-aware D2D in reuse mode

Figure 7-4 shows the signaling scheme proposed to support network controlled context-aware D2D communication with cellular uplink resource being reused. The service authorization process is not the main topic for this work and therefore it is not considered. As can been seen from this figure, several cellular users are served by the base station already with dedicated resource.

Dashed lines are used there to represent cellular uplink transmission sessions. More detailed descriptions to Figure 7-4 are provided below.

1. Users to exploit D2D service will receive the sidelink system information blocks (S-SIBs) which are broadcasted by BS and dedicated to D2D operation. In these S-SIBs, information are provided to support D2D discovery and communication processes, i.e. configuration information of resource pools used for D2D link receiving and transmitting [3GPP17-36331].
2. One D2D discoverer UE and one D2D discoveree UE form a D2D pair by exploiting either ProSe direct discovery model A or Prose direct discovery model B, as defined in 3GPP [3GPP16-23303]. In model A, one D2D device transmits a discovery message by announcing “I am here”. And in model B, a discovery message “who is there?” or “are you there?” will be transmitted. In this message, application information and reference signal for channel estimation are also conveyed.
3. A D2D discovery receiver with the discovery message successfully received will check whether it is permitted to form a D2D link with the transmitter and announce a response message to indicate its availability. If an acknowledgement message is transmitted back to D2D discoverer, certain context information will also be conveyed in the message, i.e. D2D discoveree user position and velocity and the measured channel gain information for both D2D link and cellular link (in between serving BS and the D2D discoveree).
4. D2D discoverer will send the service request to the serving BS together with certain context information, i.e. channel gain information of both D2D and cellular links, QoS of the D2D link, user positions and velocities of both D2D ends, service priority. The D2D channel measurement is performed in step 2 and the cellular links channel measurement is performed in the same way as in legacy LTE-A system by measuring the cell specific reference signals (CSRS).
5. BS will analyze the gathered context information and perform its RRM algorithm.
6. If results from the RRM algorithm admits corresponding D2D link to re-use certain frequency band of other uplink cellular users, a resource configuration message will be sent by BS to the D2D ends in DCI. Meanwhile, the cellular link with the same resource may also need to be reconfigured, in order to overcome impacts from D2D link.
7. After receiving the configuration information, D2D link will configure its transmission with the allocated resource.
8. If interference for the D2D link is sensed to be under the threshold, D2D transmission starts.

Please note that, in case of a bi-directional D2D transmission, different resource can be configured for different directions. Moreover, the D2D discovery procedure happened in step 2 and step 3 does not require D2D ends to enter RRC connected state. However, in step 4, only one D2D end needs to enter RRC connected state and send the D2D service request message. After sending the service request message, no matter whether D2D transmission is uni-directional or bi-directional, both the D2D transmitter and receiver need to listen to the DCI where D2D configuration information will be provided. Compared with the DCI used for legacy LTE-A network, the DCI for D2D operation is intended for one D2D pair instead of only one cellular UE.

In case if one D2D end is out of cellular coverage, the same procedure applies here. However, the D2D user located in cell coverage will send the D2D service request to the serving BS. After receiving the DCI, this user will forward the corresponding information to the other user who is out of cell coverage.

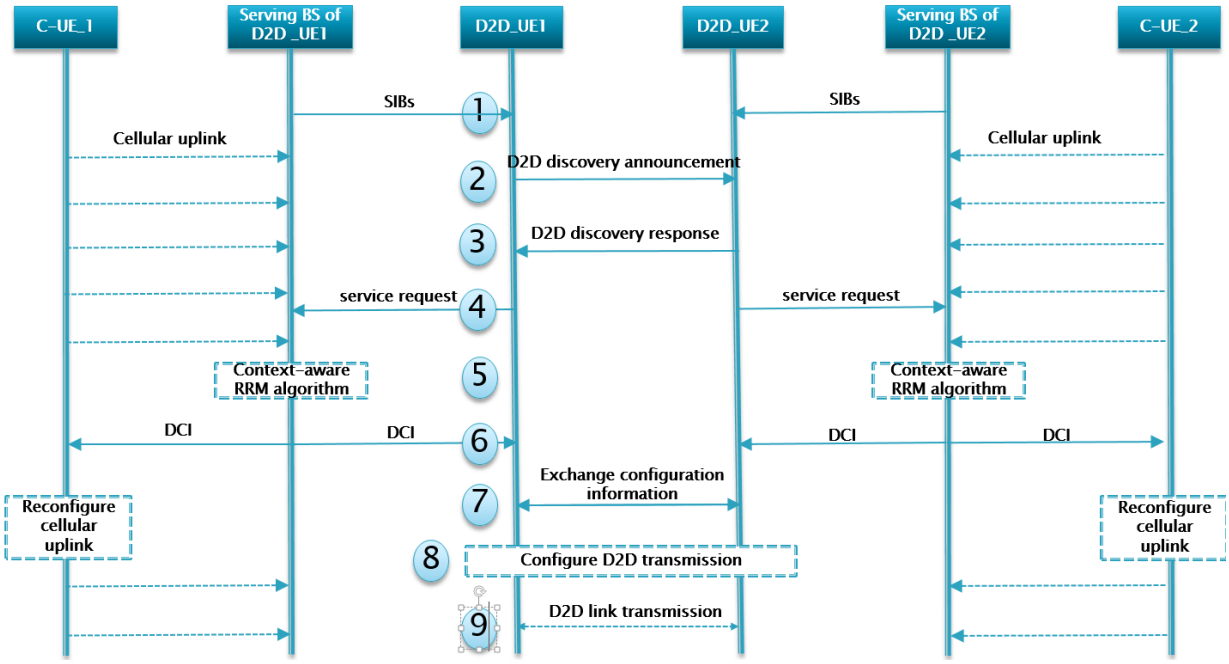


Figure 7-5 Signaling scheme in multi cells to enable context-aware D2D in reuse mode

Figure 7-5 shows the proposed solution to enable one D2D link in a scenario where two D2D users are connected to two different BSs.

1. Users to exploit D2D service will receive the S-SIBs which are broadcasted by their serving BSs and dedicated to D2D operation. In 3GPP, current defined S-SIBs will not only provide the information of resource pools used for transmission of D2D discovery and communication message in the cell, but also provide the information of resource pools used in neighbouring cells.
2. One D2D discoverer UE and one D2D discoveree UE form a D2D pair by exploiting either ProSe direct discovery model A or Prose direct discovery model B. Reference signal for channel estimation is also transmitted in this process. The D2D discovery announcement message is sent over a resource indicated in the S-SIBs from the serving BS of the D2D discoverer.
3. Since the S-SIBs received from BS#2 contain also the information of the resource pools used for D2D discovery and communication transmission in BS#1, the D2D discoveree can monitor the resource over which the discovery message was sent in previous step. The D2D discoveree will reply to the D2D discoverer, together with certain context information if the D2D discovery announcement is accepted, i.e. D2D discoveree user position and the

measured channel gain information for both D2D link and cellular links (from both the serving BS and also the neighbouring cell).

4. When a discovery acknowledgement message is detected by D2D discoverer, context information of the discoverer UE is transmitted towards the discoveree UE.
5. Both discoverer and discoveree UEs send the service request with context information to their serving BSs.
6. Both BS will analyse its gathered context information and perform its RRM algorithm.
7. For each of the BSs, if the D2D UE located in its cell coverage is allowed to transmit on certain frequency band of the uplink cellular users, a resource configuration message will be sent by the BS to the corresponding D2D end in DCI. The message contains the information about which resource should be used for the D2D end to transmit D2D data. Meanwhile, the cellular link with the same resource may also need to be reconfigured.
8. After receiving the configuration information from their serving BSs, two D2D ends will exchange their configuration information. Thus, both of them can be aware of the resource which they should listen to and receive D2D data stream from.
9. Based on the DCI acquired from step 7, each of D2D ends configure its transmission with the allocated resource. Based on the configuration information acquired from step 8, each of D2D ends configure its receiving with the corresponding resource.
10. If interference for the D2D link is sensed to be under the threshold, bi-directional D2D transmission starts.

7.3.3 Evaluation and numerical results

In order to evaluate system performance of the proposed D2D communication scheme, a system level simulator is conducted to inspect on the system capacity in the Madrid grid environment. In this scenario, the performance of outdoor users is inspected. Moreover, the number of D2D links is statistically the same as the number of cellular UEs in uplink and full buffer traffic is used for both the cellular and D2D links. In order to capture the radio propagation characters in a urban dense scenario, a 3D channel model proposed in [MET13-D61] is applied to model both the cellular and D2D links. More detailed information regarding the simulation models and assumptions are captured in Annex A.5.2.

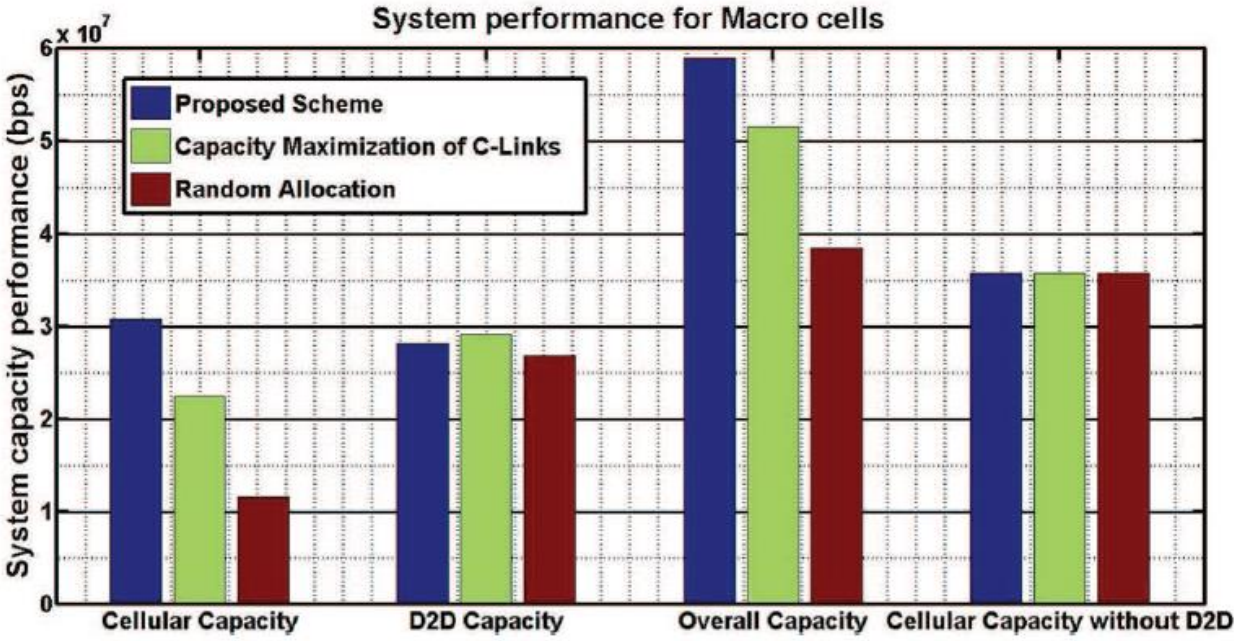


Figure 7-6 System Performance with only Macro cells (scheme No.1 of D2D power control)

We inspect on system performance of the proposed D2D communication with only macro station being deployed. The target SNR values for open loop power control are randomly generated in intervals [10dB, 15dB] for cellular links and [7dB, 12dB] for D2D links. In Figure 7-6 , notation "Cellular Capacity" represents capacity of all macro cellular links, "D2D Capacity" represents capacity of all D2D links in reuse mode, and “Overall Capacity" represents the overall capacity of both cellular and D2D links. As a comparison, the cellular capacity without D2D communication being used is also shown, with a notation of "Cellular Capacity without D2D". Moreover, the performance of another scheme denoted as "Capacity Maximization of C-Links" is also plotted where capacity of cellular links are maximized by using the algorithm proposed in [JKK+14]. Last but not least, a random allocation of cellular resource to D2D links is also inspected and shown in this work, denoted as "Random Allocation". It is to be noted that all the resources of the cellular links are reused by D2D links in the schemes "Capacity Maximization of C-Links" and "Random Allocation". However, the reuse of cellular resource is determined by the radio conditions of both cellular and D2D links in our proposed scheme.

As it can be seen from Figure 7-6 due to the extra interference from D2D links in reuse mode, the capacity of cellular links in our proposed scheme has approximately decreased 16% compared with the case where no resource of cellular links is reused. However, thanks to the contribution from D2D links, the overall capacity has an increase of 60% compared with legacy cellular network where no D2D communication is allowed.

7.4 Cooperative transmission

Cooperative D2D communications where D2D pairs implement relay functionalities to facilitate transmission between a cellular user (CU) and its base-station (BS) is a way to improve spectrum efficiency. In such scenarios there is unicast D2D communication and/or one-to-many/all D2D communication among pairs of devices over PC5* interface i.e. one of these devices can be source (DT or D2D Transmitter) while other as destination (DR or D2D Receiver). Cooperative communication scheme enables 5G RAN to dynamically allow cooperative D2D mode selection and communication, at the same time ensure interference mitigation e.g. in case of simultaneous D2D communication and CU to BS communication over the shared radio resources, etc. To enable cooperative D2D communications, among others, approaches for cooperative mode selection, relay selection, cooperative transmission and resource allocation are discussed in METIS II deliverable D6.1 [MII16-D61].

In D2D communication, interference management is one of the key topics for ensuring high spectral efficiency, and various techniques involving MIMO signal processing, power control, and transmission mode selection have been proposed to reduce the interference between the D2D pair and the cellular user (CU) or base-station (BS), especially when multiple D2D pairs are allowed to share the same channel. Some mechanisms need to be designed to further mitigate the interference both among D2D pairs and between the D2D transmitters (DTs) and the cellular system.

By allowing cooperation among DTs, as illustrated in Figure 7-7, more D2D pairs can be allowed to transmit simultaneously in the same and limited spectrum resource, increasing the spatial spectrum utilization of the system.

Through cooperation, transmitting terminals can together form a virtual antenna array to increase their transmit reliability or throughput. However, this typically requires data sharing transmission and coordinative joint transmission among cooperating terminals, which can be costly, especially for long-term resource balancing consideration. Besides, fairness consideration is also important to allow each D2D pair to achieve at least the same performance as that with no cooperation.

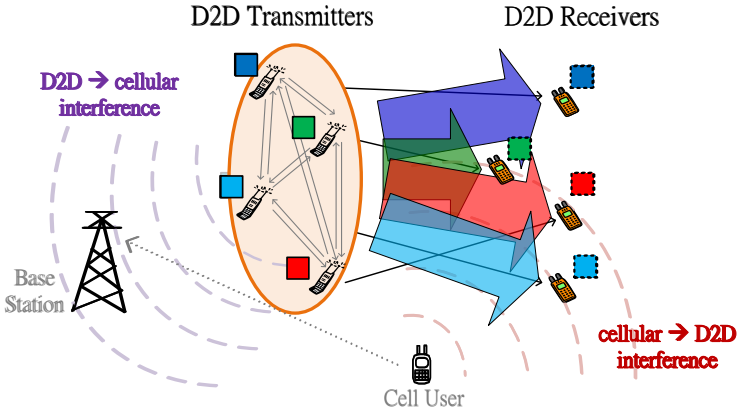


Figure 7-7 An illustration of cooperative D2D transmission

In order to increase the number of simultaneously communicating D2D pairs in the same and limited spectrum resource that the system can accommodate and to enhance the spatial spectrum utilization, we consider a two-phase cooperative D2D transmission method that consists of data-sharing transmission in phase 1 and cooperative joint transmission in phase 2. Resource balancing between the two phases is considered in our design. As illustrated in Figure 7-8, we consider a time slotted system with slot duration normalized to 1, η_1 portion of time is allocated to each DT in phase 1 and η_2 is allocated to the entire phase 2. Hence, we have $K\eta_1 + \eta_2 = 1$.

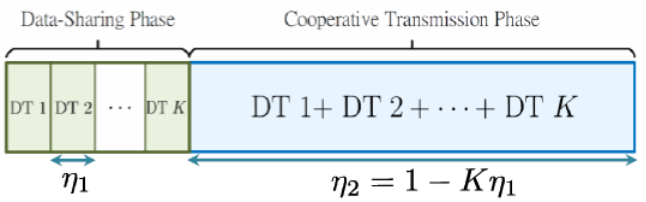


Figure 7-8: Transmission length in time of the two phase cooperation

In the data-sharing phase, each DT utilizes a physical layer multicasting scheme to efficiently transmit its data to all other cooperating DTs; and, in the cooperative joint transmission phase, the DTs adopt a multi-user block-diagonalization precoding scheme to transmit all DTs data jointly to the corresponding D2D receivers (DRs). The rate is limited by the minimum of the two phases, and the sum power over the two phases must satisfy each DTs individual long-term power constraint. More importantly, fairness must be ensured in terms of allowing each D2D pair to achieve at least the same performance as that with no cooperation. When the D2D system underlays the cellular system, the interference that the D2D causes on the cellular system must also be limited. These issues are taken into consideration in our design.

Our main objective is to find suitable covariance matrices of the designed pre-coders in phase 1 and phase 2 respectively. This is to maximize the long-term sum utility of the system, subject to long-term individual power and rate-gain constraints as well as a constraint on the interference at the BS. Here the long-term individual power constraint is used to limit the energy consumption of each device; the long-term rate-gain constraint is used to ensure that each D2D pair achieves a larger rate through cooperation; and the interference constraint limits the interference that the D2D transmission may cause on the cellular system. Formulation details of the optimization problem and the simplified implementation method to solve the optimization problem is provided in Annex A.6. Additionally, 5G RAN needs to implement below functionalities in order to enable cooperative D2D communication among devices:

- eNB provides assistance information to select D2D pairs in a particular cooperative D2D group, including determining the number of D2D pairs in the group. Based on the geo-location information of the D2D pairs, eNB can determine which D2D pairs can be grouped to perform cooperative D2D method and inform to these D2D pairs. Besides, the throughput improvement for the cooperative D2D method will be saturated when the number of D2D pairs in the group is large enough. Hence, eNB should limit the number of

D2D pairs in the group based on some information from the D2D pairs, e.g. the indication of the achievable throughput for the cooperative D2D group.

- eNB provides assistance information to perform D2D synchronization for the D2D pairs selected in the cooperative D2D group.
- eNB provides assistance information to allocate resources, e.g. channel state information (CSI) both among D2D pairs and between the DTs and the CU to the DTs, so as to perform the cooperative D2D method.

7.4.1 Evaluation of Cooperative Transmission Approaches

Performance comparisons of average sum throughput between proposed cooperative D2D transmission method and the non-cooperative method at different D2D pair number are evaluated. We consider a scenario, as shown in Figure 7-9, where the BS is located at the origin, and a CU is randomly distributed in a circular cell of radius r1 meters. And, DTs are randomly distributed with uniform distribution in a circular area with radius d_1 meters, and DRs are randomly distributed similarly in a ring with inner and outer radius equal to d_2 and d_3 meters, respectively. The simulation parameters are shown in Table 1 in the Annex A.6.

In Figure 7-10, we show the average sum rate versus SNR of the individual pairs for the case where $d_1 = 1$ meter, $d_2 = 30$ meters, $d_3 = 40$ meters and r1=300 meter for both the proposed cooperative D2D transmission method and the non-cooperative method. This is the case where the DTs are close to each other. The performance of the proposed algorithm iscompared with the non-cooperative case. From the simulation results, we can see that the average sum rate increases with the number of D2D pairs regardless of whether cooperation exists because more D2D pairs are considered for data transmission. However, the improvement will be saturated when the number of D2D pairs is large enough. Besides, the average sum rate in proposed cooperative D2D transmission method is better than that in non-cooperative method because resource balancing and fairness (rate-gain constraint) are considered in the proposed method.

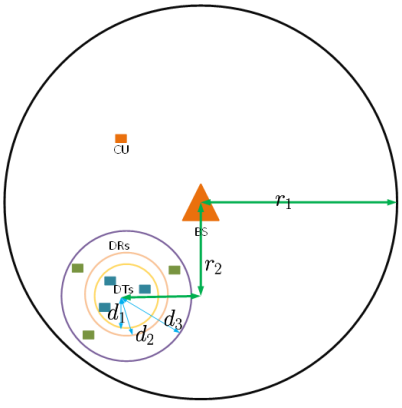


Figure 7-9: Illustrations of distributions of D2D pairs, CU and BS in the simulation

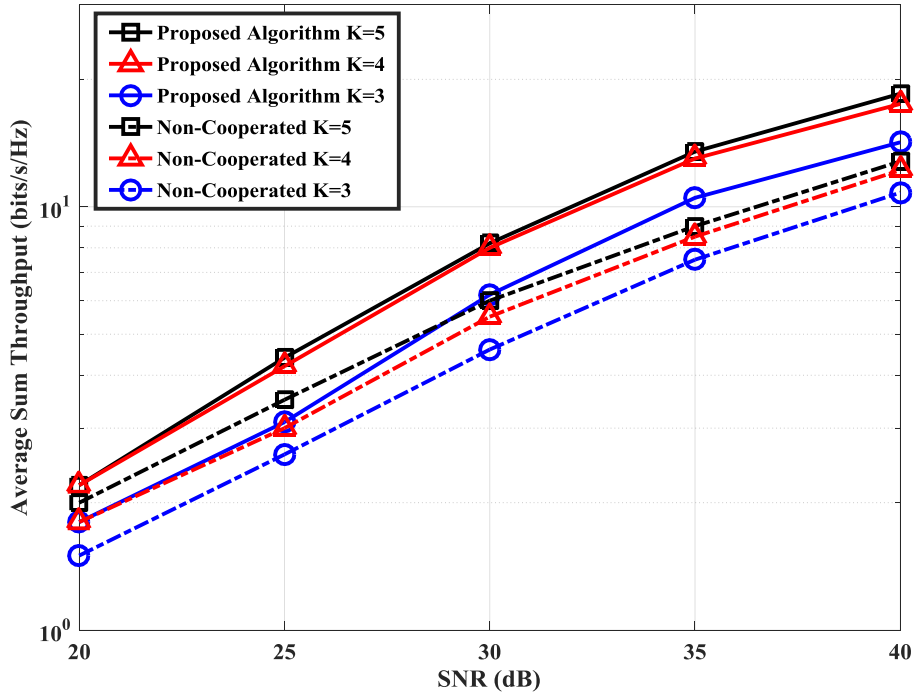


Figure 7-10: Performance comparisons between proposed cooperative D2D transmission method and the non-cooperative method at different D2D pair number

7.4.2 Cooperative Group Transmission

The concept described herein is a cooperative communication concept called Group Transmission (GT). Using GT is a way to implement joint transmission to increase the coverage and user bit rate compared to legacy single UE transmission. The operation of the GT concept is illustrated in Figure 7-11. On a high level, the first step is to discover potential group members willing to participate in the group. This is done via D2D signaling between UEs. In this step, one of the UEs is selected as the group coordinator. The next step is to synchronize UEs that are outside NW coverage by forwarding information or radio resource management information from group member UEs that are inside coverage and capable of relaying such information from an eNB. Then, when a UE within the group has UL data to transmit, it distributes the data within the group (possibly relaying may be used). In the final step, the data is transmitted jointly by all UEs in the group. Related synchronous control functions details are elaborated in [MII17-D52].

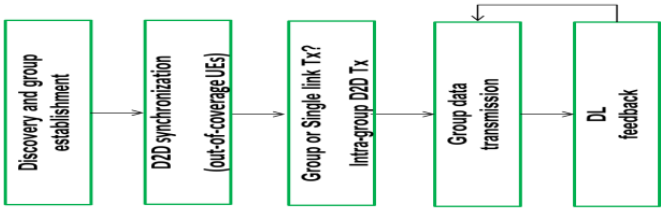


Figure 7-11: Operation of the Group Transmission concept

7.5 D2D mobility

Mobility management is one of the key components of 5G D2D. Follow-up from [MII16-D61], D2D mobility management concept are further elaborated and also evaluated here. The number of D2D/V2V devices participating and/or part of a particular group can be vary depending on the particular application scenario. For example, a platoon (of vehicles) might consist of a leader and several followers, whereas an IoT group might have significantly more communicating devices than a platoon. In the simple scenario of D2D communication, a group consists of two devices with a transmitting UE and a receiving UE. The mobility issue arises when such moving group of UEs reaches the cell edge and group members may or may not fully satisfy handover condition simultaneously and thus may not handover to the target eNB simultaneously. The current mobility scheme does not specify the procedures to handover a D2D/V2V devices in a particular group simultaneously. Besides, D2D communication over PC5 interface uses radio resources allocated by the source eNB for a particular D2D UE-pair or group. Thus, the established D2D and/or V2V link within the group or between a pair of D2D devices would be interrupted, which leads to packet loss. Moreover, each UE in the D2D and/or V2V is likely to be handed over to the target cell in an individual fashion, which leads to extra signaling overhead.

The D2D/V2V mobility management scheme, as shown in the below Figure 7-12 is designed based on the procedure of a general handover scheme. The overall procedure can be divided into five parts: Signaling quality measurement, D2D/V2V UE context management, handover decision, D2D/V2V UE context retrieval, and handover triggering. Each part will be explicitly described in the following paragraphs.

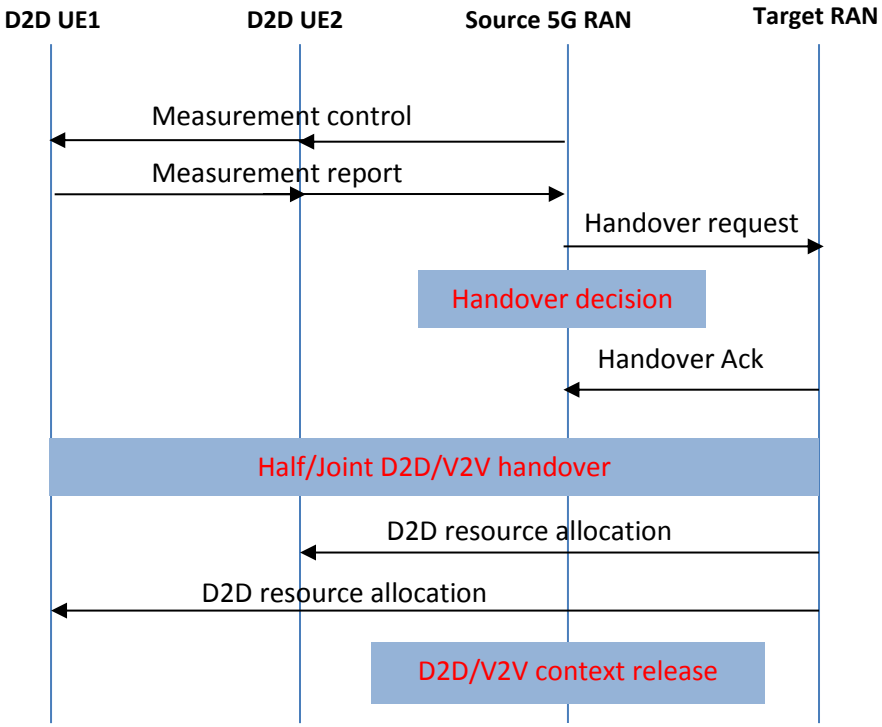


Figure 7-12: General procedures of the D2D/V2V mobility management scheme

First of all, the measurement control and measurement report messages are exchanged between the source 5G RAN and the D2D/V2V UEs. The measurement control message should include the D2D/V2V measurement ID and D2D/V2V measurement object, which identifies the RANs and D2D/V2V links to be measured by the D2D/V2V communicating UEs. To ensure inter-RAT D2D/V2V compatibility and service continuity, the measurement targets could include intra-PLMN 5G RAN, inter-PLMN 5G RAN and other RANs. The measurement target information can be sent to the UEs via dedicated signaling for connected state UEs or on-demand SIB for idle state UEs. On the other hand, the quality of the D2D/V2V links within the D2D/V2V group needs to be inspected. UEs should measure the D2D/V2V reference signal over the 5G PC5* interface. To allow identification of D2D/V2V UEs across multiple RANs/AIVs, 5G RAN should also use unified and global D2D/V2V ID for each individual D2D/V2V UE.

The obtained reference signal received powers (RSRP) between the UEs and the source 5G RAN, RSRP between the UEs and the target RANs, the D2D/V2V link quality of the D2D/V2V group and other context information will then be sent to the source 5G RAN for handover decision making. The measurement report can be either periodic or event triggered depending on the 5G RAN configuration.

The D2D/V2V mobility management scheme adopts the “make-before-break” rationale. This analogy ensures that the D2D/V2V context of the group is transferred from the source 5G RAN to the target RAN prior to the actual handover execution, so that target 5G RAN can prepare customized D2D/V2V resources and quick connection/configuration setup. Due to broad range of

D2D/V2V applications, the D2D/V2V groups from different applications may have distinct service requirements, i.e. D2D/V2V topology, D2D/V2V traffic types and D2D/V2V communication types. A unified service cannot fulfill the requirements of various types of D2D/V2V applications.

To ensure that the requirements of different D2D/V2V applications can be satisfied, a D2D/V2V mobility management function may be used for storing and analyzing the D2D contexts. This function can be implemented in the 5G core network, 5G RAN, or the D2D/V2V UE itself. The D2D/V2V mobility management function is able to store and manage D2D/V2V-related information for mobility management. For example, by analyzing the D2D/V2V UE traces and locations, the D2D/V2V mobility management function should be able to generate suitable D2D/V2V context of the group.

With a proper multi-AIV signal measurement and D2D/V2V signal measurement, handover decisions can be smartly made to ensure D2D/V2V service continuity. Taking a pair of two UEs under D2D/V2V communication as an example, the source 5G RAN decides whether to perform joint or half handover based on the signal quality of the measured links. In general, joint handover is performed if both UEs satisfy a handover condition to the same target RAN. On the other hand, half handover is performed if only one of the UEs satisfies handover condition to the target RAN and has bad connection with the source 5G RAN while the other UE still has good RSRP with the source RAN.

After determining the type of D2D/V2V handover to be performed, the source 5G RAN sends a handover request to the target RAN. As mentioned earlier, the target RAN could be either intra-PLMN 5G RAN, inter-PLMN 5G RAN or other AIV RAN. Therefore, similar to X2 interface in LTE, an X2* interface is needed in 5G network to allow coordination of different RATs. The handover request message should contain the result of D2D/V2V handover decision and the D2D/V2V context of the D2D/V2V pair. The target RAN is ready to prepare new D2D/V2V resources after receiving the handover request

The actual handover execution begins when the source 5G RAN receives the handover request acknowledgement from the target RAN. In the case of a joint handover, both UEs would detach from the source 5G RAN and synchronize to the target RAN simultaneously. The D2D/V2V resources from the target RAN is allocated to the pair via downlink message. The D2D/V2V pair can start using the new D2D/V2V resource directly after completion of joint handover. In the case of half handover, the front UE (the UE which is closer to or in better coverage to the target RAN) should synchronize to the target RAN upon receiving the half handover command. However, the rear UE (the UE which is farther away or at the edge or poor coverage to the target RAN) should store the half handover command and synchronizes to the target RAN only when the RSRP between itself and the target RAN satisfies the criteria of half handover. In the case of half handover, the D2D/V2V pair keeps using the D2D/V2V resource from the source 5G RAN until the both UEs successfully connect to the target RAN. As such, the D2D/V2V service is maintained even during the period when one is waiting for another.

Evaluation of D2D Mobility Management Approaches

Simulations were performed to evaluate the performance of the proposed D2D mobility management scheme. The D2D reliability is used as the performance metrics. D2D reliability is calculated as the number of successfully received D2D packets over the number of transmitted D2D packets. Comparisons are made between the proposed D2D mobility management scheme and the conventional LTE handover scheme. Figure 7-13 shows the D2D reliability with respect to D2D device velocities. The handover delay of each individual UE is kept at either 2ms or 200ms during the simulation.

Figure 7-13 shows that the D2D reliability decreases as the velocity increases for both the LTE handover scheme and the D2D mobility management scheme. However, the D2D mobility management scheme helps to achieve significantly higher D2D reliability than the LTE handover scheme. The main reason is that for LTE handover scheme, the ongoing D2D communications is terminated when any of the D2D UE satisfies handover condition. The D2D service will not be resumed until both UEs successfully handover to the same target cell and re-establish the D2D link. Another observation is that when the handover delay of each individual UE is short (2ms), there is almost no drop in D2D reliability regardless of the velocity and inter-device distance. However, if the delay is high (200ms), the inter-device distance and velocity have more influence on the D2D reliability. The reason is that half handover has twice the handover delay of joint handover, due to the fact that the two UEs in the pair do not handover simultaneously in the half handover procedure. However, such an influence is not obvious if the simulated handover delay of an individual UE is low (The difference between 2ms and 4ms is small). In general, Figure 7-13 shows that the D2D/V2V mobility management scheme is capable of handling vehicles or devices at high speed.

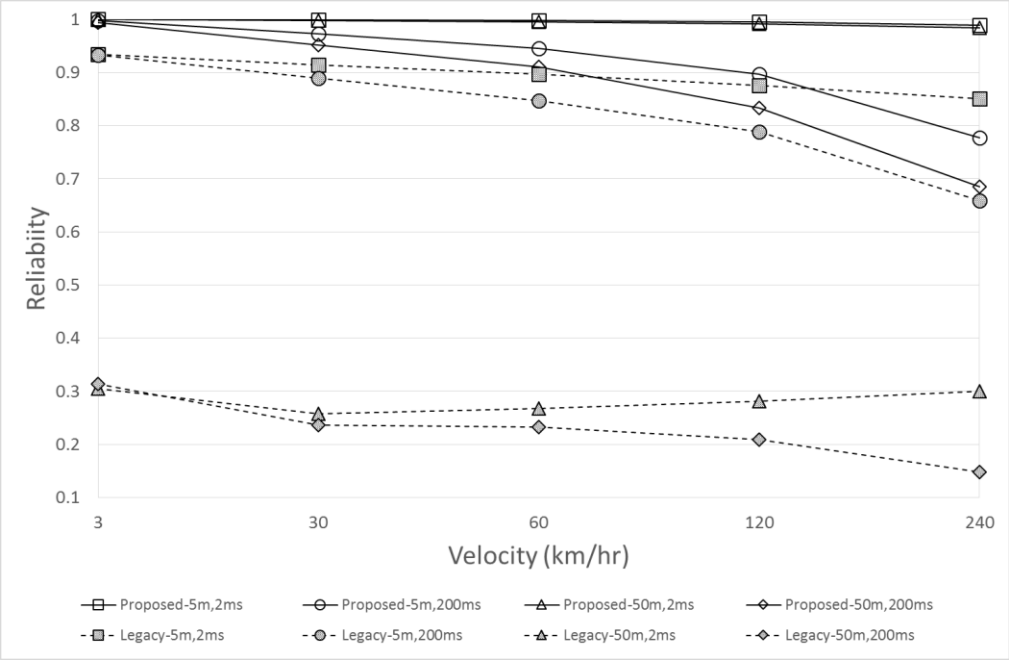


Figure 7-13: The effect of velocity on D2D reliability

7.6 Context Aware Group Mobility Management

One aspect which could benefit from D2D communication or communication via secondary interfaces is the location management; in such approaches the existing grouping schemes and secondary interfaces/connections are used for handling certain control operations. Solutions that are available in the literature consider grouping for minimizing the mobility management cost. Such solutions exploit the same mobility pattern of the devices and only the group leader device performs mobility management on behalf of the devices in the same group. However, they have some very important drawbacks, since often they are applicable only in cases with cyclic movements or in environments with limited dynamicity in terms of mobility. This problem is even more important in uMTC (URLLC) scenarios with tight latency requirements.

A solution could be to take advantage of already available groups formed for other purposes (device tracking, car to car clustering, etc.) and perform group location update where one device is responsible for informing the network on behalf of the group and having per user (or per user group) Tracking Area List (TAL), and not predefined ones. For forming the groups and for exchanging information in the group D2D communication or secondary interfaces (e.g., 802.11p, 802.15, etc.) may be used.

Figure 7-14 shows the TAU process for the proposed mechanism. As it is shown only the group representative performs TAU whereas all the other devices once they join a group they stop all the location management processes. It should be highlighted that once a device/UE joins a group then the network is being informed about the fact that this device is associated with the group and that the group representative performs TAU on behalf of the device/UE. The cost for the formation and the maintenance of a group should be considered and the tradeoff among the gains of the TAU suppression and the group formation and maintenance should be investigated. However, the proposed scheme with the reduced user oriented TALs enables the significant reduction of the paging area without potential paging misses/failures.

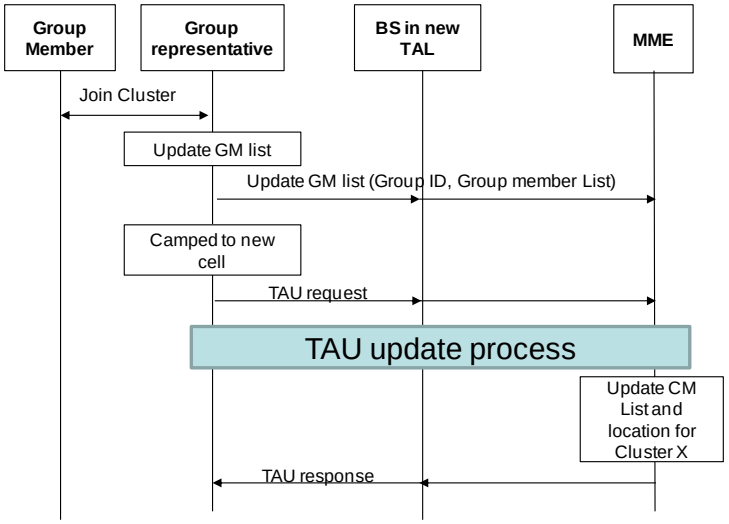


Figure 7-14: TAU update in the group location management by exploiting multi connectivity

The proposed scheme allows for a significant reduction of the location management overhead. For this analysis we compare to two solutions in terms of location management cost and the paging accuracy/cost. The one is presented in [FLY+14] and is based again on group location management but the grouping is performed by the core network (e.g., MME) after certain number of joint TAU requests – this mechanism from now on will be called Group Mobility Management scheme or GMM. The other is an extension of this scheme for increased accuracy. According to the former scheme, the devices perform periodically TAU and once the network identifies a stable pattern it proceeds in grouping and it notifies the devices that they should stop their TAU when they enter a new TAL and only the cluster head performs TAU. The devices still have to perform the periodic TAU so as to ensure the stability of the cluster. The drawback of this scheme is that every time that the UE has an incoming call the whole TAL should be paged because it is not certain whether the device has remained in the group; this will be ensured only after the periodic TAU. Even in the case that the overall TAL is paged it may still be the case that we will have paging misses, since we may have a device that left the cluster and hasn't yet performed periodic TAU. In the case that we want to ensure that the cluster will be stable we have in case of every change in the direction of a UE to perform a TAU – something that will increase significantly the signaling cost but will ensure that there will no page misses.

On the other hand, in the proposed solution the network is informed every time that the UE performs cell reselection. This ensures that the group can be accurately located on a per cell granularity. Considering groups of 3 to 8 devices (typical for platoons [MKA+14]) and car flow of 17800 for 15 minutes [MKA+14], a cell reselection rate 30 per UE/hour [CNT07] and a TAU rate of 1.2 per UE/hour we can quantify the number of bytes transmitted in the RAN for every case [ALU13]. Figure 7-15 shows the signaling cost by using the aforementioned assumptions for small (3 cluster members) and large (8 cluster members) clusters. It can be observed that the proposed solution outperforms the GMM when a small number of re-clustering is considered. When the re-clustering number increases the GMM seems to perform better, but this is not actually true, because then the cluster is not stable and in case of a paging there will be a miss. On the other hand, if every time that we have a re-clustering need the UEs send a TAU (enhanced GMM) to keep the cluster stable then the signaling cost increases. Figure 7-16 presents the paging gains of our mechanism against the GMM mechanism (and LTE and Enhanced GMM) where there is the need to page the whole TAL.

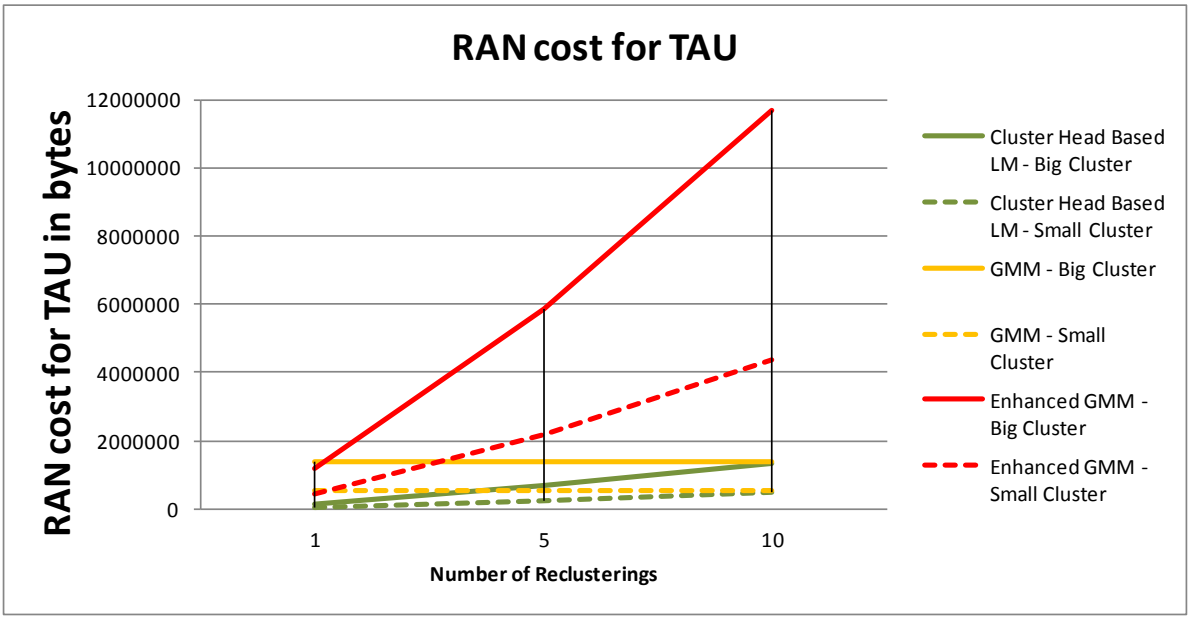


Figure 7-15 RAN cost for TAU for the three different schemes considering reclusterings for small and larger clusters

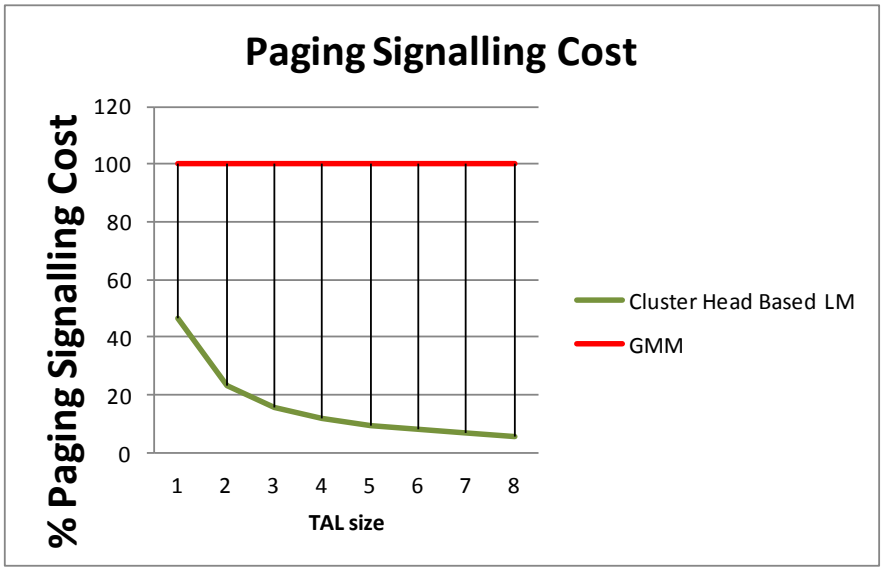


Figure 7-16 Paging signaling cost

8 Conclusions

The METIS-II project aims at developing a comprehensive and detailed 5G RAN design in order to support the standardization. This deliverable describes the final view of the asynchronous Control Plane (CP) functions. The deliverable can be said to consist of three different parts:

- **Overall CP framework and architecture** such as CP/UP split options, different radio configurations for slicing and SON functions
- **Other network CP functions related to both Core Network (CN) and RAN** such as security, QoS, packet switching between CN and RAN, spectrum management and integration with Wifi
- The **CP functions** such initial access, RRC state handling, mobility and native support of D2D

The two first bullets mostly addresses Task 6.2 while the last bullet addresses Task 6.1 (see section 1.1, Objective of the document).

The sum of all the CP functions and concepts in this deliverable are designed to fulfill a wide variety of requirements such as future-proof design, high energy efficiency, support of beamforming mobility and higher connection reliability.

In general, it is assumed that all CP functions herein are common for all AIVs regardless the carrier frequency used, except possibly if slicing is used where some CP functions can be turned off. Different options on how to best split the CP or UP for different physical architecture options are further analyzed in this deliverable. This also includes using a split option at the PDCP layer (with PDCP located in the central unit). The PDCP layer is assumed be used as an aggregation/split layer for LTE and NR tight integration. The tight integration between LTE and NR is an important aspect of 5G already from the beginning to enable a smooth migration and higher reliability of the new 5G networks.

The control plane signalling between RAN and CN can be even further reduced by using Ethernet datagrams instead of GTP tunnels because each datagram directly contains the address information needed to forward the packet. The spectrum allocation should be dynamic in both time and space based on traffic demand, different type of services, as well as the different radio access technologies (e.g. LTE and NR).

One design goal for METIS-II has been to “push down” functionalities to the RAN that formerly needed CN – RAN signaling in order to reduce the CN-RAN signaling and improve efficiency. One way to do this is to enable a more RAN based paging. This concept also benefits from a new middle UE state between IDLE and ACTIVE, which allows the network to be able to page the UE more accurately (involving less NBs). The new state also allows for faster transition to CONNECTED mode, i.e. the UE can start transmitting faster compared to LTE due to less required signaling . These concepts have been proposed by METIS-II at an early stage and are now part of the ongoing 3GPP NR work.

This deliverable also describes several concepts for the initial access and how the network control signaling related to the initial access (system information) shall be designed. The RACH concept developed in this deliverable enables both better capacity and allows smoother prioritizing between users with different delay requirements compared to LTE. Further on, the SI shall be designed so that it is not being broadcasted periodically but instead is on-demand mode. Together with self-contained RS (with the data) and longer time between synchronization signals. The 5G system may sustainably decrease the energy consumption compared to LTE.

To fulfill thehigh reliability requirements of 5G, the mobility functionality is of vital importance. The mobility concept of in 5G is enhanced by several new methods including UE autonomous mobility, make-before-break and mobility concepts for URLLC. Make-Before-Break is thoroughly investigated in this deliverable, together with considerations for multi-connectivity for a centralized RAN architecture.

D2D is expected to be an integral component of 5G systems and natively supported in the CP protocol stacks of 5G. This deliverable addresses several D2D components and their CP design perspective to enable this. One of these components is the use of context awareness for D2D, and it is shown in this deliverable that it can increase the overall capacity. Another component is the cooperative relaying, which can increase the UL user throughput.

A summary of the key takeaways for each chapter and the studies in each chapter is given below.

	Study	Key takeaway
Control Plane Framework	**Control plane framework**	All CP functions are common for all AIVs, except possibly if slicing is used.
	Centralization/Distribution of CP/UP network functions	Analyses how the different functional split options can be best mapped to different deployment scenarios (physical architectures)
	LTE-NR tight integration architecture options	PDCP used for aggregation/split layer for LTE and NR tight integration
	5G RAN Configuration Modes	Identification of potential differences and similarities among the RAN configurations for the various services. Link of the RAN configurations with the end-to-end slicing concept.
	SON architecture and Control/Management Plane	Improving radio network creation, design and optimization processes and enabling the integration in End-to-End service oriented orchestration architecture through the definition of an “abstraction level” for RAN domain.
CN and RAN	**Security**	5G security should support the new Connected Inactive state and enable an efficient security solution for mMTC
	Connectionless packet switching in and between	The control plane signaling between RAN and Core can be strongly reduced by using Ethernet datagrams instead of GTP

	the NextGen RAN and Core	tunnels because each datagram directly contains the address information needed to forward the packet.
	QoS framework for multiple air interfaces	Open up for more flexible 5G QoS
	Integration with WiFi	When WiFi access points are co-located with 3GPP sites, tight integration provides large gains. If not, network assistance by sending average load information of each system to UEs is a good solution compared to a blind upper layer integration.
	Asynchronous spectrum management	Efficient use of the spectrum considering the dynamic (in time and space) traffic demand, the different type of services, as well as the different radio access technologies (e.g. LTE and NR)
State handling	**State Handling**	The novel RRC state model allows all kinds of services supported by the same RRC state machine.
	RRC Connected State	The RRC Connected state allows extreme data rates and network controlled mobility similar to LTE
	RRC Connected Inactive State	New RRC Connected Inactive state improves UE battery consumption and allows lower CP latency by faster transition to CONNECTED mode, i.e. the UE can start transmitting faster now compared to LTE due to less signaling is required. RRC Connected Inactive can be used as the main power saving state for the UE.
	RRC Idle State	The RRC Idle state in 5G is used for example during UE power-on, inter-RAT cell selections, fault recovery. In Idle state the 5G UEs can be paged from Core network.
	Inter-RAT state transitions with RRC Connected Inactive	New RRC Connected Inactive state can be used for inter-RAT cell reselections between 5G and LTE when connected to NextGen Core with harmonized RAN based tracking areas allowing fast system access and data transmission.
Initial Access	**Group based RACH**	Group based initial access for reducing the collision rate. Instead of grouping the devices on a per mobility case, we also consider the service requirements and we prioritize the devices with stringent delay requirements
	URLLC	Combine preambles transmissions from certain UEs to prioritize service requests in case of collisions.
	Paging	The RAN based paging concept benefits from the new Connected Inactive which allows the network be able to page the UE more accurately (fewer NBs broadcasting the paging).
	5G RAN lean design	With on-demand SI, avoiding always on reference symbols (self-contained with data) and longer time between

		synchronization signals, the 5G system may sustainably decrease the energy consumption compared to LTE.
Mobility	**Mobility**	The Mobility concept of in 5G is enhanced by several new methods including UE autonomous mobility, make-before-break and mobility concepts for URLLC. Make-Before-Break is thoroughly investigated in this deliverable, together with considerations for multi-connectivity for a centralized RAN architecture.
	Mobility and Multi Connectivity in centralized RAN architecture	Centralization of the UE AS context, RRM, and multi-connectivity has a potential to improve the efficiency of various INACTIVE and ACTIVE state mobility and multi-connectivity procedures.
	Inter-RAT mobility	Inter-RAT mobility between LTE and 5G allows seamless inter-RAT mobility and multi-connectivity by anchoring the RAN/CN interface to the LTE or 5G network node. Inter-RAT mobility can support UEs in both RRC Light Connected in LTE and RRC Connected Inactive in 5G state thus allowing low energy dissipation and fast state transition to Connected state from either of the systems in inter-working scenarios.
	UE aspects of mobility	The proposed CSI adds for UE related mobility measurements are shown as sufficient for the RSRP accuracy measured with SSS in some scenarios. CSI-RS is only required in certain scenarios as common RRM measurement for both IDLE and CONNECTED mode. The evolution of UE capability signaling for LTE-5G tight-interworking scenarios were evaluated and forward looking options are presented for efficiently implementing UE capabilities signaling for LTE and 5G.
	Context aware mobility]	Context information are used to enhance the accuracy of mobility prediction that will enable a uniform service experience for the user, even in deep shadow regions or coverage holes.
	Data Analytics for Traffic Engineering	Use of UE profiles built based on past user behavior for predicting the future network requirements. This approach significantly reduces the signaling cost for context information transfer
D2D	**Context-aware D2D communication to serve mMTC**	Application of D2D communication can efficiently enhance the system performance of mMTC services in cellular network, in terms of power consumption and service availability.
	Context-aware D2D underlay to improve system capacity	Taking into account of different context information, D2D communication can reuse the resource of cellular users for its transmission and therefore improve the network capacity.

Cooperative transmission	Cooperative D2D communication as proposed here, along with proposed interference management, significantly improves spectrum efficiency.
D2D mobility	Mobility management scheme is proposed to fulfill the D2D service continuity requirements ensuring that UEs in the D2D/V2V group (of two or more) can be handed over to the target RAN without breaking the D2D/V2V link, also enabling URLLC.
Context Aware Group Mobility Management	Use of context information to perform group mobility management. This significantly reduces the UE tracking and tracking area update cost.

9 References

[3GPP09-32101] 3GPP TS 32.101, “Telecommunication management; Principles and high level requirements (Release 9)”, September 2009.

[3GPP13-R1132030] 3GPP Tdoc R1-132030, “Channel models for D2D performance evaluation”, Ericsson, ST-Ericsson, May 2013.

[3GPP15- 36211] 3GPP TS 36.211, Evolved Universal Terrestrial Radio Access (E-UTRA); Physical channels and modulation

[3GPP15-36912] 3GPP TR 36.912, “Feasibility study for further advancement for E-UTRAN (LTE Advanced) (Release 13)”, December 2015.

[3GPP15-45820] 3GPP TR 45.820, “Cellular system support for ultra-low complexity and low throughput Internet of Things (CIoT) (Release 13)”, November 2015.

[3GPP16-22891] 3GPP TR 22.891 Technical Specification Group Services and System Aspects; Feasibility Study on New Services and Markets Technology Enablers; Stage 1 (Release 14), 2016

[3GPP16-23303] 3GPP TR 23.303, “Proximity-based services (Release 14)”, December 2016.

[3GPP16-23799] 3GPP TR 23.799, “Study on Architecture for Next Generation System (Release 14)”, December 2016.

[3GPP16-R1163961] 3GPP Tdoc R1-163961, “Final Report of 3GPP TSG RAN WG1 #84bis”, ETSI, April 2016.

[3GPP16-R2167708] 3GPP Tdoc R2-167708, “Paging and location tracking in RRC_INACTIVE”, Nokia, Alcatel-Lucent Shanghai Bell, November 2016.

[3GPP16-R2168858] 3GPP Tdoc R2-168858, “Text Proposal to TR 38.804 on on-demand SI provisioning for NR”, NTT DOCOMO, November 2016.

[3GPP16-R3162728] 3GPP Tdoc R3-162728, “Flexibility of RAN functional split”, Nokia, Alcatel-Lucent Shanghai Bell, KT, November 2016.

[3GPP16-RP160671] 3GPP Tdoc RP-160671, “3GPP Working Item Description: New SID Proposal: Study on NR New Radio Access Technology”, NTT DOCOMO, March 2017.

[3GPP17-36300] 3GPP TS 36.300, “E-UTRAN; Overall Description; Stage 2; (Release 12)”, Mar. 2017.

[3GPP17-36331] 3GPP TS 36.331, “Radio Resource Control (RRC); Protocol specification (Release 14)”, March 2017.

[3GPP17-38912] 3GPP TR 38.912, “Study on New Radio (NR) Access Technology” (Release 14)”, March 2017.

[3GPP17-38804] 3GPP TR 38.804, “Study on New Radio Access Technology Radio Interface Protocol Aspects (Release 14)”, February 2017.

[3GPP17-38913] 3GPP TR 38.913, “Study on Scenarios and Requirements for Next Generation Access Technologies”, Release 14)”, March 2017.

[3GPP17-R11700031] 3GPP Tdoc R1-1700031, “Coexistence of NR DL and LTE”, Huawei, HiSilicon, January 2017.

[3GPP17-R11700841] 3GPP Tdoc R1-1700841, “NR LTE Coexistence”, Qualcomm, January 2017.

[3GPP17-R21700060] 3GPP Tdoc R2-1700060, “Implications of High Frequency Bands on Mobility”, Nokia, Alcatel-Lucent Shanghai Bell, January 2017.

[3GPP17-R21700122] 3GPP Tdoc R2-1700122, “Fast and Early Radio Measurement Filtering for Multi-Connectivity”, Nokia, Alcatel-Lucent Shanghai Bell, January 2017.

[5GN17-D32] ICT-671584 5G NORMA: Deliverable D3.2 “5G NORMA network architecture – Intermediate report”, January 2017.

[5GPPP16] 5G PPP White Paper, “5G PPP use cases and performance evaluation Models”, April 2016.

[ALU13] Alcatel Lucent Bell Labs, Application Note, “Managing the Signalling Traffic in Packet Core”, 2013.

[CISCO+16] Cisco Visual Networking Index: Forecast and Methodology, 2015–2020, June 2016

[CKS+15] K. Chatzikokolakis, A. Kaloxylos, P. Spapis, et al., “On the Way to Massive Access in 5G: Challenges and Solutions for Massive Machine Communications”, EAI International Conference on Cognitive Radio Oriented Wireless Networks (CrownCom), April 2015.

[CNT07] A. Catovic, A.; M. Narang, A. Taha, “Impact of SIB Scheduling on the Standby Battery Life of Mobile Devices in UMTS”, Mobile and Wireless Communications Summit, July 2007.

[CPRI13] “Common Public Radio Interface (CPRI); Interface Specification”, V6.0, August 2013

[CROSSHAUL] 5G PPP project 5G-Crosshaul, http://5g-crosshaul.eu/.

[CWH+13] F. Chang, H. Wang, S. Hu, and S. Kao, “An Efficient Handover Mechanism by Adopting Direction Prediction and Adaptive Time-To-Trigger in LTE Networks”, Computational Science and Its Applications – ICCSA 2013, Lecture Notes in Computer Science, Volume 7975, pp. 270-280, 2013.

[DDL15] B. Debaillie, C. Desset, F. Louagie, “A Flexible and Future-Proof Power Model for Cellular Base Stations,” IEEE Vehicular Technology Conference (VTC-Spring), Glasgow, Scotland, UK, 2015.

[DRW+09] K. Doppler, M. Rinne, C. Wijting, C. B. Ribeiro, K. Hugl, "Device-to-device communication as an underlay to LTE-advanced networks," IEEE Communications Magazine, vol. 47, no. 12, pp. 42-49, December 2009.

[ERI11] M. Ericson, “Total Network Base Station Energy Cost vs. Deployment”, ,” IEEE Vehicular Technology Conference (VTC-Spring), Budapest, Hungary, 2011

[FLY+14] H. Fu, P. Lin, H. Yue, G. Huang, C. Lee, “Group Mobility Management for Large-scale Machine-to-Machine Mobile Networking,” IEEE Transactions on Vehicular Technology, vol. 63, pp. 1296-1305, March 2014.

[GSMA+17] GSMA; https://www.gsmaintelligence.com/, 2017-04-03

[GZ16] J. Gebert, D. Zeller, ”Fat pipes for user plane tunnelling in 5G“, IEEE Conference on Standards for Communications & Networking (CSCN), Berlin, October 2016.

[ITU09] ITU-R Report M.2135, “Guidelines for evaluation of radio interface technologies for IMT-Advanced”, December 2009.

[JKK+14] J. Lianghai, A. Klein, N. P. Kuruvatti, and H. D. Schotten, “System Capacity Optimization Algorithm for D2D Underlay Operation”, Workshop on 5G Technologies at IEEE International Conference on Communications (ICC), Sydney, Australia, June 2014.

[KKP+13] Y. Kim, H. Ko, S. Pack, W. Lee, and X. Shen, “Mobility-Aware Call Admission Control Algorithm With Handoff Queue in Mobile Hotspots”, IEEE Transactions on Vehicular Technology, vol. 62, no. 8, October 2013.

[KKS+13] N. P. Kuruvatti, A. Klein, J. Schneider, and H. D. Schotten, “Exploiting Diurnal User Mobility for Predicting Cell Transitions”, IEEE Global Communications Conference (GLOBECOM) WS, Atlanta, USA, 2013.

[KKS15] N. P. Kuruvatti, A. Klein, and H. D. Schotten, “Prediction of Dynamic Crowd Formation in Cellular Networks for Activating Small Cells”, IEEE Vehicular Technology Conference (VTC-Spring) Workshop, Glasgow, Scotland, 2015.

[LNM+13] M. Lauridsen, L. Noël, and P. Mogensen, “Empirical LTE Smartphone Power Model with DRX Operation for System Level Simulations”, IEEE Vehicular Technology Conference (VTC-Fall), 2013.

[MBS+10] R. Madan, J. Borran, A. Sampath, N. Bhushan, A. Khandekar, and J. Tingfang, "Cell Association and Interference Coordination in Heterogeneous LTE-A Cellular Networks", IEEE Journal on Selected Areas in Communications, 2010.

[MCC+11] J. Márquez-Barja, C.T. Calafate, J.-C. Cano, and P. Manzoni, “An overview of vertical handover techniques: Algorithms, protocols and tools”, Computer Communications, vol. 34, no. 8, pp. 985-997, ISSN 0140-3664, 2011.

[MET13-D11] ICT-317669 METIS, Deliverable D1.1, Version 1, “Scenarios, requirements and KPIs for 5G mobile and wireless system”, April 2013.

[MET13-D61] ICT-317669 METIS, Deliverable D6.1, Version 1, “Simulation guidelines”, October 2013.

[MET15-D15] ICT-317669 METIS, Deliverable D1.5, Version 1, “Updated scenarios, requirements and KPIs for 5G mobile and wireless system with recommendations for future investigations”, May 2015.

[MET15-D64] ICT-317669 METIS, Deliverable D6.4, Version 2, “Final report on architecture”, January 2015.

[MII16-D11] ICT-671680 METIS-II, Deliverable D1.1, Version 1, “Refined scenarios and requirements, consolidated use cases, and qualitative techno-economic feasibility assessment”, January 2016.

[MII16-D22] ICT-671680 METIS-II, Deliverable D2.2, Version 1, “Draft Overall 5G RAN Design”, June 2016.

[MII16-D51] ICT-671680 METIS-II, Deliverable D5.1, Version 1, “Draft Synchronous Control Functions and Resource Abstraction Considerations”, May 2016.

[MII16-D61] ICT-671680 METIS-II, Deliverable D6.1, Version 1, “Draft Asynchronous Control Functions and Overall Control Plane Design”, June 2016.

[MII17-D23] ICT-671680 METIS-II, Deliverable D2.3, Version 1, “Performance evaluation results”, February 2017.

[MII17-D42] ICT-671680 METIS-II, Deliverable D4.2, Version 1, “Final air interface harmonization and user plane design”, April 2017.

[MII17-D52] ICT-671680 METIS-II, Deliverable D5.2, Version 1, “Final Considerations on Synchronous Control Functions and Agile Resource Management for 5G”, March 2017.

[MKA+14] M. R. Karim, A. A. SaifizuI, R. Syahira, and H. Yamanaka, “The effect of Gross Vehicle Weight on Platoon Speed and Size characteristics on Two-Lane Road”, Intemational Conference on Innovative Trends in Multidisciplinary Academic Research (ITMAR), October 2014.

[MKS+10] A. Merentitis, A. Kaloxylos, M. Stamatelatos, and N. Alonistioti, "Optimal periodic radio sensing and low energy reasoning for cognitive devices," IEEE Mediterranean Electrotechnical Conference (MELECON), April 2010.

[NGMN14] NGMN Alliance, Deliverable D2, “Recommended Practices for Multi-vendor SON Deployment”, January 2014.

[NGMN15] NGMN Alliance, “5G White Paper”, February 2015, available at http://www.ngmn.org/fileadmin/ngmn/content/images/news/ngmn_news

[NOK14] Nokia Networks, White Paper “Looking Ahead to 5G: Building a virtual zero latency gigabit experience”, 2014.

[NOK16] Nokia Networks, White Paper “Multi-Layer and Cloud-Ready Radio Evolution towards 5G”, 2016.

[ORI14] ETSI GS ORI 002-1/2, “Open Radio equipment Interface (ORI); ORI interface Specification; Part 1: Low Layers / Part 2: Control and Management”, V4.1.1, October 2014.

[PWH13] K. Pentikousis, Y. Wang, and W. Hu, “Mobileflow: Toward software-defined mobile networks”, IEEE Communications Magazine, vol. 51, no. 7, pp. 44-53, July 2013.

[QXY+10] T. Qu, D. Xiao, D. Yang, W. Jin, and Y. He., "Cell selection analysis in outdoor heterogeneous networks", IEEE International Conference on Advanced Computer Theory and Engineering (ICACTE), 2010.

[SMA+16] I. Da Silva, et al., "Impact of network slicing on 5G Radio Access Networks," European Conference on Networks and Communications (EuCNC), Athens, 2016.

[SMS+16] I. Da Silva, G. Mildh, M. Säily, and S. Hailu, “A Novel State Model for 5G Radio Access Networks”, IEEE International Conference on Communications (ICC) Workshop, 2016.

[SRG+12] P. Spapis, R. Razavi, S. Georgoulas, Z. Altman, R. Combes, and A. Bantouna, "On the role of learning in autonomic network management: The UniverSelf project approach," Future Network & Mobile Summit, 2012.

[TFA+16] S. Tombaz, P. Frenger, F. Athley, et al., “Energy Performance of 5G-NX Wireless Access Utilizing Massive Beamforming and an Ultra-lean System Design”, IEEE Global Communications Conference (GLOBECOM), San Diego, CA, USA, December 2015.

[XPM+14] D. Xenakis, N. Passas, L. Merakos, and C. Verikoukis, “Mobility Management for Femtocells in LTE-Advanced: Key Aspects and Survey of Handover Decision Algorithms”, IEEE Communications Surveys & Tutorials, 1st Quarter 2014.

[ZN13] K. Zhou, N. Nikaein, “Packet aggregation for machine type communications in LTE with random access channel,” IEEE Wireless Communications and Networking Conference (WCNC), April 2013.

A Annex

A.1 Functional split options within the RAN

The following figure gives an overview about different functional split options within the UP of the radio protocol stack ("horizontal split") denoted here as M1 – M8 [MII16-D22]. In addition, it also shows the separation between NFs of CP and UP ("vertical split") and the corresponding interfaces in between (see the red arrows marked by (1) – (12)). More details with respect to the impact of horizontal splits can be found in METIS-II Deliverable D4.2 [MII17-D42].

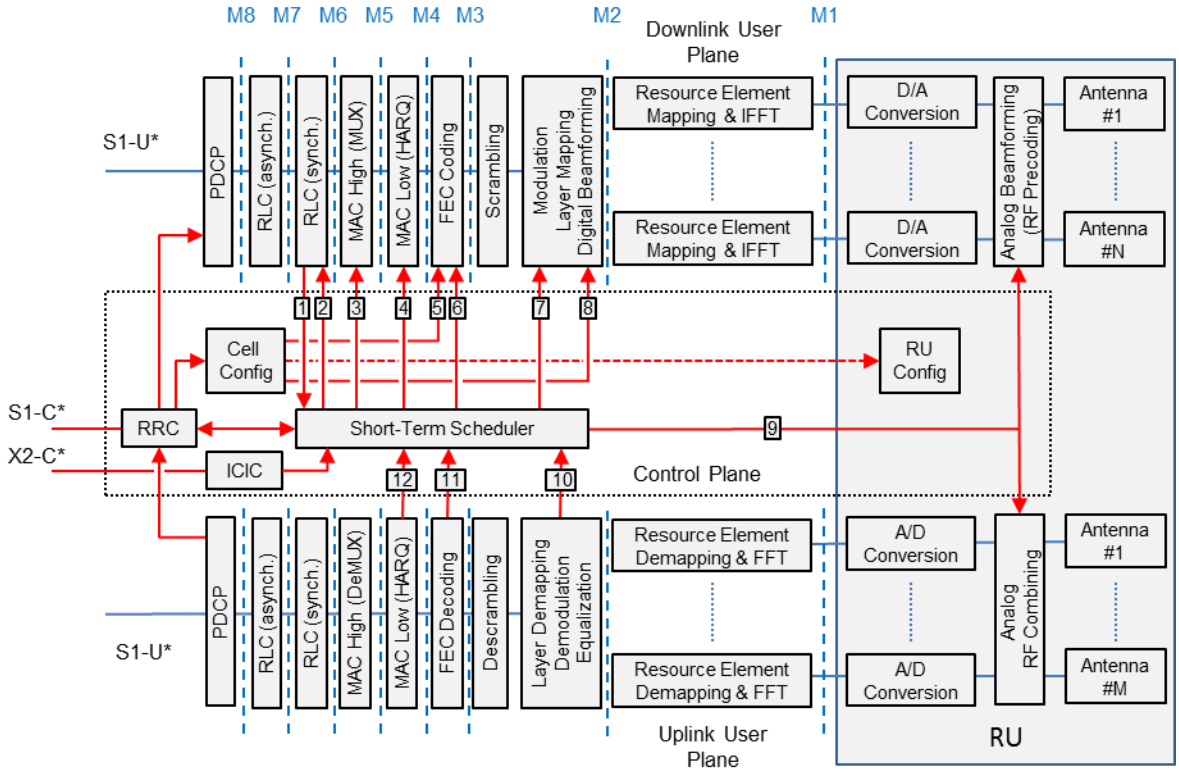


Figure A-1: RAN functional split options (horizontal/vertical splits)

A.2 Initial Analysis of UCs for Slicing

In order to identify the expected differences in terms of functionalities and parameterizations in the protocol stack of the base station we need to examine closely the characteristics and the requirements of the foreseen use cases. For this we have analyzed the METIS use cases so as to identify potential commonalities them, and group them according to their characteristics. In METIS project 25 UCs have been identified – a subset of them has been described with less detail than the others which have been presented thoroughly in [MET13-D11] and [MET15-D15]. Based on the identified similarities we have concluded in 9 groups of UCs presented in Table A-1.

Table A-1: Use Case grouping of METIS I according to the requirements analysis

UC group	UC Characteristics	METIS UC
UC Group 1	Static users Small density Very High Throughput No Energy Efficiency Requirement	UC1, UC13, UC15
UC Group 2	Limited mobility Medium density High throughput No Energy Efficiency Requirement	UC2, UC3, UC6, UC7, UC14
UC Group 3	Very high mobility Medium density High throughput No Energy Efficiency Requirement	UC8
UC Group 4	Limited mobility Medium density High throughput No Energy Efficiency Requirement	UC4, UC9
UC Group 5	Limited mobility High density Low Throughput High Energy Efficiency Requirement	UC11, UC20
UC Group 6	No mobility Low Density Low Throughput No Energy Efficiency Requirement Very high reliability Very low latency	UC5
UC Group 7	Very high mobility Medium density Low Throughput No Energy Efficiency Requirement Very high reliability Very low latency	UC12
UC Group 8	No mobility Low Density Low Throughput High Energy Efficiency Requirement Very high reliability Very high latency	UC10
UC Group 9	No mobility Low Density High throughput No Energy Efficiency Requirement Very high reliability Very low latency	UC17, UC21

The previous analysis implies that the METIS UC, even though that they had diverse characteristics and were user oriented, focuses on the user experience. This implies that the characteristics of each UC description mainly focus on the user (including both humans and machines) requirements. The user requirements tend to be related with the UP but they may also

affect the CP. This observation leads to the outcome that the UP and the CP should be approached separately for each UC.

Additionally, for the CP and the UP the analysis should focus on the special characteristics of each UC that affects the protocol stack and not all the characteristics. This will further facilitate the proper analysis of the protocol stacks and the required diversity for each different UCs.

A.3 Lean design

This section gives an overview of the gains with a 5G system using more lean design than current LTE system. Since the work in 3GPP is just started it is impossible to understand in detail what the lean design aspects will look like. Therefore, this study utilizes an approximate method to illustrate the effect of a more lean design. Table A-2 shows the CP overhead for LTE and 5G for inactive (zero users) and active cells

This study (either or both) investigates the lean design on a principle basis for an UDN network using a variable of higher probability for micro sleep of the gNB than for LTE due to lean design.

Table A-2 Control plane overhead for LTE and 5G, inactive (zero users) and active cells.

Signaling	LTE		NR	
	Inactive cell	**Active**	**Inactive cell**	**Active**
Reference Symbols	On (CRS)	On (CRS, DMRS etc)	Off	On
PDCCH	Off	On	Off	On but more efficient than LTE
System info (MIB + SIBs)	On	On	MIB only	On
Synch signals	On	On	On	On
Total CP overhead	20%	28%	10%	28%
Probability to micro sleep	Low due to CRS, 16% [TFA+16]	0%	High, 71% [TFA+16]	0%

There is no beam-forming in this study. Assuming BF may even further increase the energy efficiency for NR since some signals can be transmitted only via the connected beam.

We assume an asynchronous network, i.e. CP data and UP may interfere with each other.

A.3.1 Methodology and parameters

The evaluation was made by using a theoretical calculation of the user bit-rate and cell throughout. The average SINR is calculated as a function of m users per cell as:

$$SINR_i(m) = \frac{P_i \cdot G_i}{\sum_{m \neq i}^{m} P_m \cdot G_m + N_0} \approx \frac{G_i}{\sum_{m \neq i}^{m} G_m} \approx \frac{1}{F_{const}(m)}$$

Thus, the total interference from adjacent cells in LTE/5G is modeled as the inter-cell interference factor F_{const} (basically the inverse of the geometry factor) assuming equal transmit power from all cells. We assume an interference limited scenario and can thus neglect the noise impact. Note that this means we calculate the SINR as if all users in the cell were placed on the same spot; in a real simulation the users would be spread out and experience different inter-cell interference factors. However, since we are just interested in an average value this is good enough for us. Also, the inter-cell interference factor F_{const} is based on the full load scenario, i.e. all cells are fully loaded and transmit over the whole bandwidth. However, we are also interested in the case where cells are not loaded (the zero user case). Therefore, the inter-cell interference factor is modified based on the probability of zero users in a cell as:

$$F(m) = (1 - P(x = 0, m)) \cdot F_{const}$$

Where $P()$ is a Poisson distributed random function and F_{const} the constant inter cell interference factor if there is a full load in all cells.

However, to model the interference from the control signaling we expand our original equation to:

$$SINR_i(m) = \frac{1}{F(m)} = \frac{1}{F(m) + F_{CP}}$$

The F_{CP} is the inter cell interference caused by the control plane signaling. The CP overhead F_{CP} is further divided into two parts, one when there are active users ($1\text{-}P(x=0,m)$) and one when there are zero active users ($P(x=0,m)$). The overhead for the CP is different if there are active users in a cell, denoted CP_{active}, or if there are zero users, denoted $CP_{zerousers}$, see Table A-2 and Table A-3 for the values for LTE and 5G. Thus, the inter cell interference caused by the control plane signaling, F_{CP}, can be written as

$$F_{CP} = (1 - P(x = 0, m)) \cdot F_{const} \cdot CP_{active} +$$
$$P(x = 0, m) \cdot ((1 - Pr_{celldtx}) \cdot F_{const} \cdot CP_{zerousers} + Pr_{celldtx}(F_{const} \cdot 0))$$

Where $Pr_{celldtx}$ is the probability for the cell to enter cell DTX (micro sleep) assuming there is not user in the cell. Note that if there is no user in the cell and the cell enters the cell DTX mode, there will be no interference from the CP signaling (nor the UP of course).

Finally, the SINR for each user is used to calculate the user bit-rate as:

$$BR_i = (1 - CP_{overheadratio}) \cdot BW \cdot \log_2(1 + \frac{SINR_i(m)}{ShannonLoss})$$

Where the user bit-rate BR_i is deducted with the $CP_{overheadratio}$, which is the resources used by all CP overhead.

We also need to calculate the power consumption. The model is taken from Metis R2.3 and [TFA+16].

$$P_{in\_active} = P_0 + \Delta_p \cdot P_{max} \cdot \lambda \cdot \alpha_{PSD} + P_1 \cdot \lambda$$

Where λ is the bandwidth load which is calculated here based on the average non zero active users (1 - $P(x=0,m)$) in a cell

$$P_{in\_zerousers} = P_{sleepCellDtx} \cdot P_{sleep} + \left(1 - P_{sleepCellDtx}\right)$$

Table A-3 list the parameters used to calculate the user SINR, bit rate and power consumption.

Table A-3 Parameters used for calculating user bit-rate and base station power consumption

Parameter	LTE	5G	Comment
Number of cells	Varied	Varied	
System area	Fixed	Fixed	
Cell radius	100 to 1000 m	100 to 1000 m	
Antenna	Omni	Omni	No sectors or BF
Inter cell interference factor F'	0.4	0.4	
Traffic	FTP like	FTP like	If a user is active, it takes as much BW as available
User speed	30 m/s	30 m/s	
Shannon Loss	2	2	
Bandwidth	5 MHz	5 MHz	
Control plane signaling overhead ratio	0.28	0.28	If more than zero active users
Control plane signaling overhead if zero users	0.20	0.10	If no users, some RS and signaling etc can be turned off
Probability for cell DTX if zero users	1 - 0.84	1 – 0.29	A cell that enters cell DTX sleeps has no interference from CP or UP [TFA+16]
P_0 [W]	44.7	44.7	Power consumption parameter from [MII17-D23]

Δ_P	3.5	3.5	Power consumption parameter from [MII17-D23]
P_{max} [W]	46	46	Power consumption parameter from [MII17-D23]
P_1	2.2	2.2	Power consumption parameter from [MII17-D23]
P_{sleep} [W]	27.1	27.1	Power consumption parameter from [MII17-D23]
$P_{sleepCellDtx}$	0.84	0.29	The ratio of lower power consumption in cell DTX mode [TFA+16]

In order to investigate the effect of what happens when a network is densified (i.e. smaller cells), we evaluate the performance for different cell densities, going from 1000 to 100 m. Note that for each evaluation we have a fixed number of users in the area. This means that the probability for zero users increases with *decreased* cell radius. However, smaller cells also leads to more handover events. The traffic type is a full buffer traffic, i.e. if there is an active user it will occupy the whole bandwidth of a cell. If there is more than one active user in the cell, they share the bandwidth equally. To be an active user, the user either has to transmit user data or make a handover as shown in Figure A-2. All other procedures such as user measurement, paging etc. are not taking into account explicitly (this is implicitly handled by the cell sleep DTX ability described above). The probability to transmit data is 1% and the time spent sending is set to 1 s. The number of handovers are a function of the user speed and the cell radius, assuming a simple model that the users moves straight through the cells. As can be seen in Figure A-2, the time when the user is active due to handover are dominant at small cell radius due to the higher number of handovers.

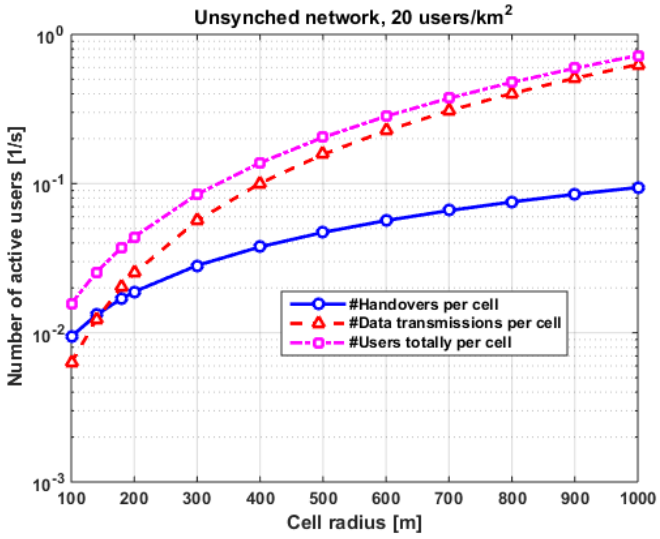


Figure A-2 Active users per cell for different cell radiuses and divided into data transmission and handover occasions. Note that the number of users are always constant over the system area.

A.3.2 Lean design user performance

Figure A-3 shows the user throughput vs. cell throughput for the case with 20 users per km^2. Left sub-figure shows the user throughput vs. the cell throughput. Each point represents a different number of cells in the system area while the number of users in the system area is constant. This means there will be fewer users the more cells in the system (see Figure A-2).

The right sub-figure of Figure A-3 shows the same thing but with the cell radius on the x-axis. Since we assume asynchronous network the 5G case can lower its interference when the cell has zero active users due to better cell DTX probabilities. Thus, there is a clear gain for the 5G compared to LTE at lower loads (and smaller cell radiuses). At high load the performance is the same for both LTE and 5G. The user throughput gain for 5G at 200m is around 40%.

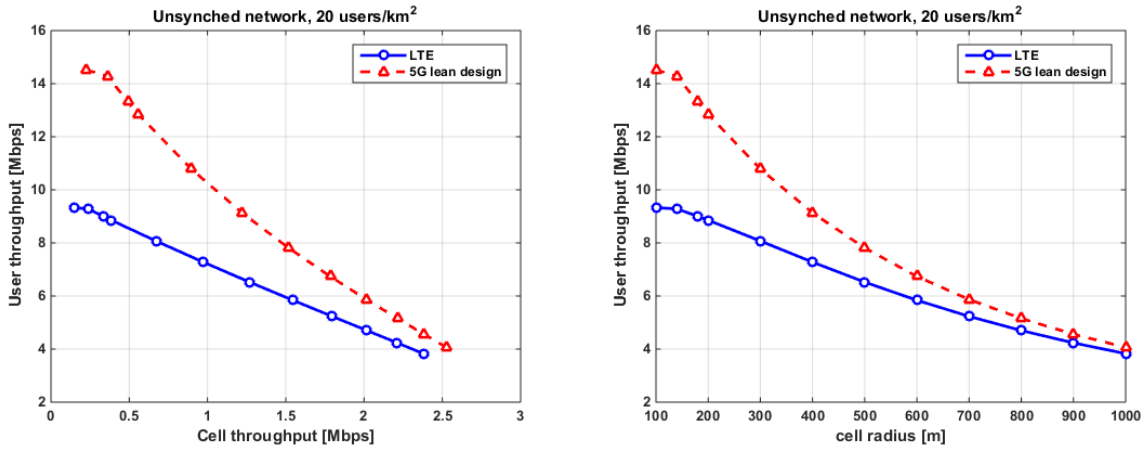


Figure A-3 User throughput vs. cell throughput for the case with 20 users per km^2. Left-figure shows the user throughput vs. the cell throughput. The right figure shows the same thing but with the cell radius on the x-axis.

A.3.3 Power consumption

Figure A-4 shows the power consumption. The left subfigure of Figure A-4 shows the total system power consumption divided by the system area (W/km^2). As can be seen, 5G power consumption is lower compared to LTE, especially for increased cell density and lower user densities. The right sub-figure shows the relative power consumption per cell for 5G over LTE. As can be seen, the power consumption is lower for the 5G case due to same reason as for the higher user bit-rate, namely 5Gs improved ability for cell DTX compared to LTE. For example, the decrease in power consumption for 5G at 200 m is around 50% when there are 5 users/km^2, 30% when there are 20 users/km^2 and 15% when there are 50 users/km^2. In general, the power savings is in the same ballpark as in [MII17-D23, Section 3.3] and [TFA+16].

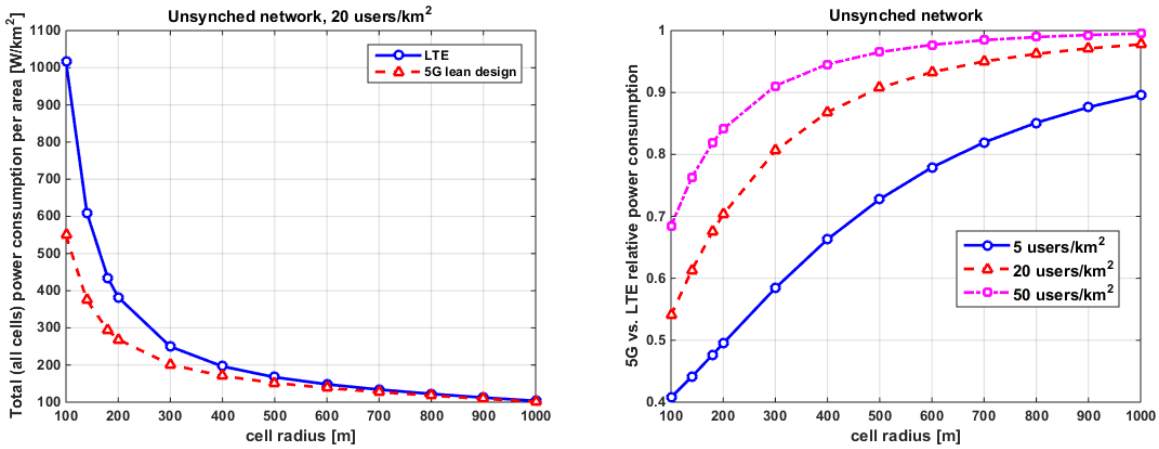


Figure A-4 Left: Cell power consumption as a function of cell radius. Right: Total system power consumption per bit. Note that there are 20 users per km² regardless of cell radius which means that the probability for zero users increases with *decreased* cell radius.

Note that we do not change the power consumption model for smaller cell radiuses, we use the same parameters all the time. In reality it might be possible to replace the base station at smaller radiuses with e.g. micro cell and potentially lower the power consumption, but this is not taken into account here.

A.3.4 Power consumption per bit

Figure A-5 , right subfigure, shows the power consumption per bit per area as a function of the cell radius for LTE and 5G. The left sub-figure shows the relative 5G power consumption per bit vs. LTE. As can be seen the 5G improvements are largest for the case with lowest user density. The decrease in power consumption per bit for 5G is now around 50% at 200 cell radius and 20 users per km2, which is a combination of the gains for higher user throughput and lower power consumption.

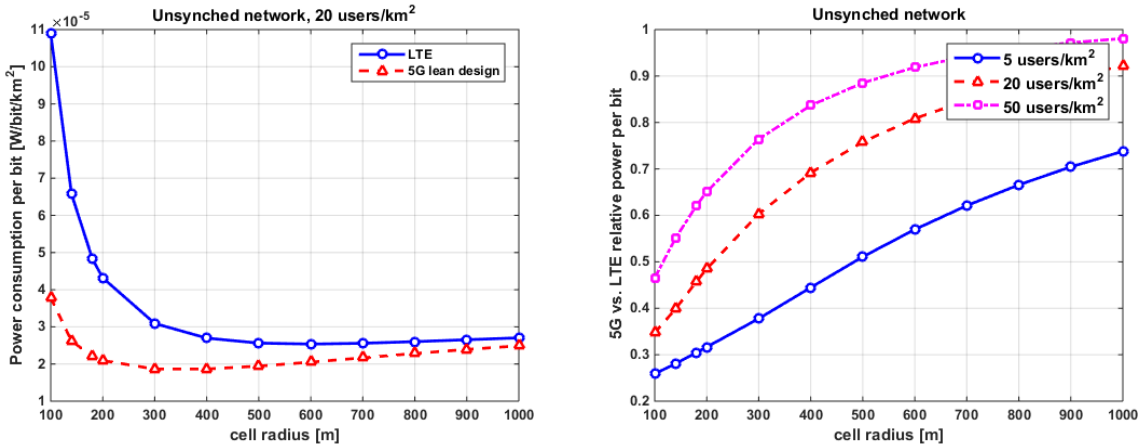


Figure A-5 Power consumption per bit. Left: Power consumption per bit per area as a function of the cell radius. Right Power consumption per bit as a function of the cell

radius for different number of users per area. Note that there are 20 users per km² regardless of cell radius which means that the probability for zero users increases with *decreased* cell radius.

A.3.5 Conclusions

This study shows the impact if the cell DTX possibilities can be improved compared to LTE. The better cell DTX possibilities of 5G will probably mainly be a combination of lower CP signaling when there is no active user in a cell and lower RS overhead due to more self-contained data transmission.

Assuming a 2.5 times higher probability to enter micro sleep for a cell (cell DTX), the energy consumption per bit for 5G is now around 50% lower at 200 m cell radius and 20 users per km2, which is a combination of the gains for higher user throughput and lower power consumption. Note that this study assumes asynchronous network, with a synchronous network the energy consumption gain for 5G will be slightly lower.

A.4 Context-aware D2D communication in mMTC

A.4.1 Radio link enablers

In this section, we discuss four procedures as potential solutions enabling context aware D2D communication for MTC services.

Initial UE grouping

Once a remote sensor device is attached to a network, initialization is required. Figure A-6: shows the procedure for initialization of remote UE as following.

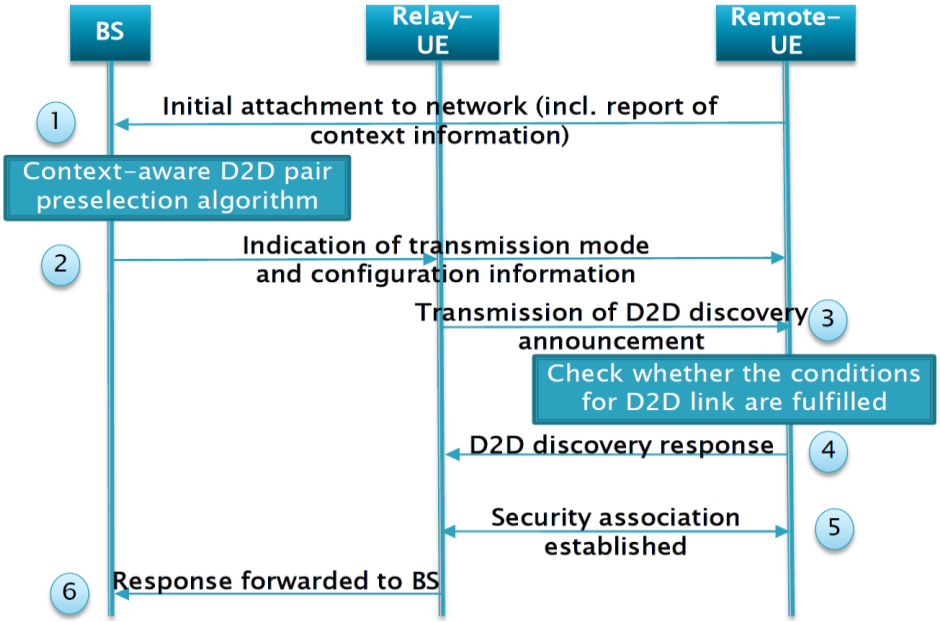


Figure A-6: Initialization procedure of a remote UE

1. The UE's location and other information are reported to its serving BS when UE is initially attached to the network. BS will analyze the location, battery level, traffic type and reference signal received power (RSRP) of connected MTC UEs to perform the context aware D2D pair preselection algorithm. This algorithm decides the proper transmission mode of the new UE, e.g. whether it is configured to report directly to BS or to report through a relay device.
2. After obtaining decision of the context-aware D2D preselection algorithm, BS will page both the relay and remote UEs and indicate their transmission mode, Meanwhile, involved UEs also check other configuration information carried by sidelink system information blocks (S-SIBs) and downlink control information (DCI), e.g. the resource used for D2D discovery and communication, transmission power control of D2D link and conditions that D2D pairs should fulfill.
3. The configured relay UE sends discovery announcement message on the specific time and frequency resource and ID of the target remote UE is included in this message. In this step, reference symbols for the purpose of D2D channel estimation will also be transmitted.
4. Based on the calculated RSRP of D2D link, remote UE determines whether the request is accepted or not (based on the information acquired from BS in step 2. And the decision is transmitted back to the relay UE.
5. If the request is accepted, a security association between D2D ends is established by exchanging messages with the help of security algorithms. In case the D2D discovery request is not accepted, a rejection message will be sent back to the relay UE.
6. The discovery acceptation/rejection message is further forwarded to the BS from relay UE. In case a rejection message is received, BS will avoid setting up the same D2D connection in future. Furthermore, the procedure from step 1 to step 5 will be repeated to select a new relay UE for the remote UE.

In this proposed signaling scheme, the D2D S-SIBs provide the less frequently changed configuration information, e.g. spectrum band used for D2D link discovery and communication. Thus, the relay UE transmits the discovery announcement message on the configured resource so that the monitoring UEs that are interested in these messages can receive and process the messages. Moreover, since both D2D ends are already synchronized to the same access node and they are statically located within the proximity of each other, a synchronization procedure between the D2D ends is not necessary. Last but not least, if a D2D pair is successfully established, certain context information should be stored at both relay and remote UEs.

Update of D2D pairs

In case if certain conditions are not fulfilled for D2D communication anymore, BS will perform the context-aware D2D preselection algorithm again and reset the D2D pair. Figure A-7: shows the procedure for updating of D2D pairs. Only two remote UEs for D2D pair are given here, in case a group of remote UEs are assigned with the same relay UE.

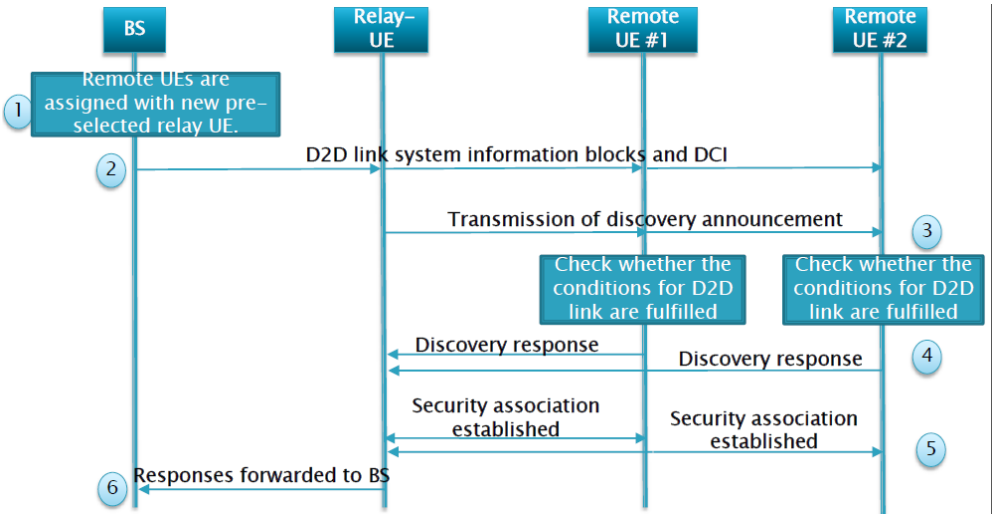


Figure A-7: Procedure of D2D cluster update

1. BS will analyze the location, battery level, traffic type and RSRP of its served MTC UEs and perform the context aware D2D pair preselection algorithm.
2. BS configures UEs with S-SIBs for direct D2D discovery, e.g. resource used for D2D discovery. UEs will also be informed by DCI for the dedicated control information, e.g. IDs of the pre-selected remote and relay UEs and conditions that D2D link pairs should fulfill.
3. Relay UEs send discovery announcement to target remote UEs, with its own ID and IDs of the target UEs being conveyed in the announcement message.
4. Based on the calculated D2D link RSRPs, target UEs determine whether the request is accepted or not (with the help of DCI acquired from BS). And the decision will be fed back to the relay UE.
5. If the request is accepted, a security association could be established by exchanging messages between UEs with security algorithms.
6. The result on D2D pair update are further transmitted from the relay UE to the serving BS. If the request is not accepted, the relay UE should inform the serving BS. Then BS reselects the relay UE as shown from step 1 to step 5 and avoids to pair the previous selected relay UE with this remote UE in future.

If the relay UE transmits a discovery announcement to a group of remote UEs in step 3, each remote UE in step 4 can respond by picking up a resource from a resource pool for D2D discovery indicated by S-SIB. Moreover, the D2D pair update procedure takes place on a large time scale, e.g. with a periodicity of one day or even larger. With this approach, extra signaling overload and device power consumption can be efficiently under control. The formed D2D group will be exploited for uplink data transmission of remote UEs before the next D2D group update procedure taking place.

D2D communication

In case if certain remote UEs are paged by network or if they have data packets to transmit in uplink, the formed D2D groups as stated in the previous subsections will be exploited to relay the data of remote UEs in uplink. Figure A-8 shows the uplink data transmission procedure. We show

here only two remote UEs for sidelink communication as an example, in case if a group of UEs are paged or they have uplink data in their buffers to transmit. Following steps provide details of this scheme.

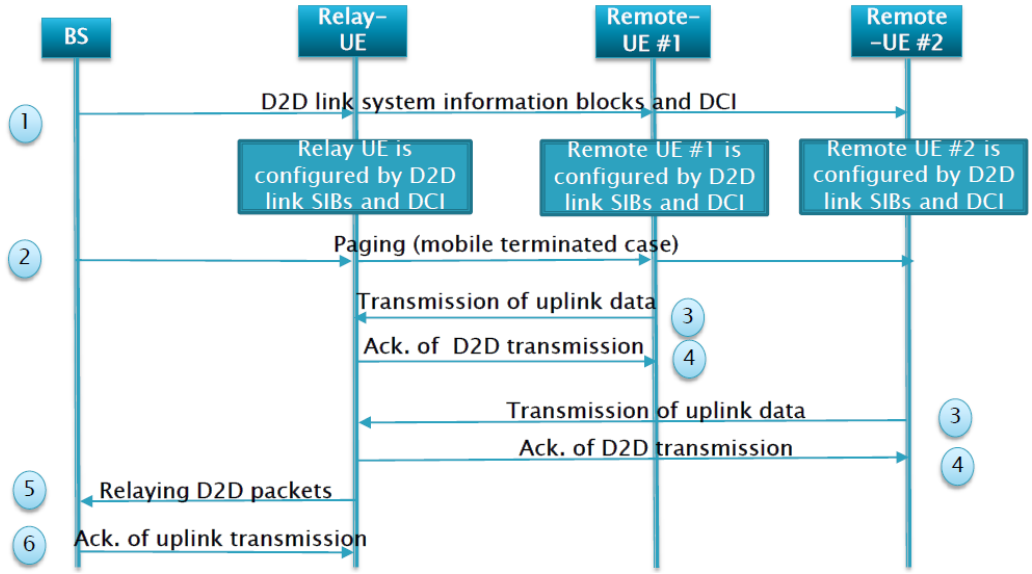


Figure A-8 D2D uplink report procedure

1. Relay and remote UEs involved in D2D communication are configured by the serving S-SIBs and DCI for D2D communication. These information are configured and stored at both relay and remote UEs when the D2D discovery procedure is accepted.
2. In mobile terminated case, one remote UE or a group of remote UEs will be paged by BS, or
3. In mobile originated case, a remote UE tries to transmit its data packet to its relay UE. This step also includes the random access procedure, D2D link connection setup procedure between the relay UE and remote UE, also D2D retransmission if an error is experienced. In this step, both D2D ends should be aware of the time and frequency resource for D2D transmission.
4. After successful receiving packets from remote UE(s), relay UE replies with an acknowledgment to the remote UE(s).
5. Upon the successful receiving of packets from remote UE(s), relay UE will forward the successfully received packets to the serving BS. This process can be performed as a normal cellular uplink transmission where a control plane (CP) connection needs to be established. Another alternative is that the received packets will be stored in relay UE and then transmitted together with the packet of the relay UE. In this case, an advantage in power saving for relay UE can be introduced, since relay UE only needs to wake up and perform the CP connection establishment procedure for one time.
6. Upon successfully receiving packets from the relay UE, BS sends an acknowledgment message to the relay UE.

In this procedure, relay UE should be informed by the network regarding whether it is allowed to compress its own uplink packets together with user packets from multiple remote UEs and send them together to BS.

D2D monitoring and release

Once a D2D group is established, both D2D ends and BS will monitor the condition for D2D pair and decide to release the D2D pair if it is necessary, e.g. due to a low battery level of relay UE or a change in propagation loss of D2D link. A D2D link can be monitored at the same time as uplink transmission procedure takes place. If one D2D end decides to release the sidelink, relay UE reports this decision to BS and both the relay and remote UEs erase the stored context information. Figure A-9 demonstrates the procedure of releasing a D2D pair, originated by remote UE. Correspondingly, details are also given in the followings.

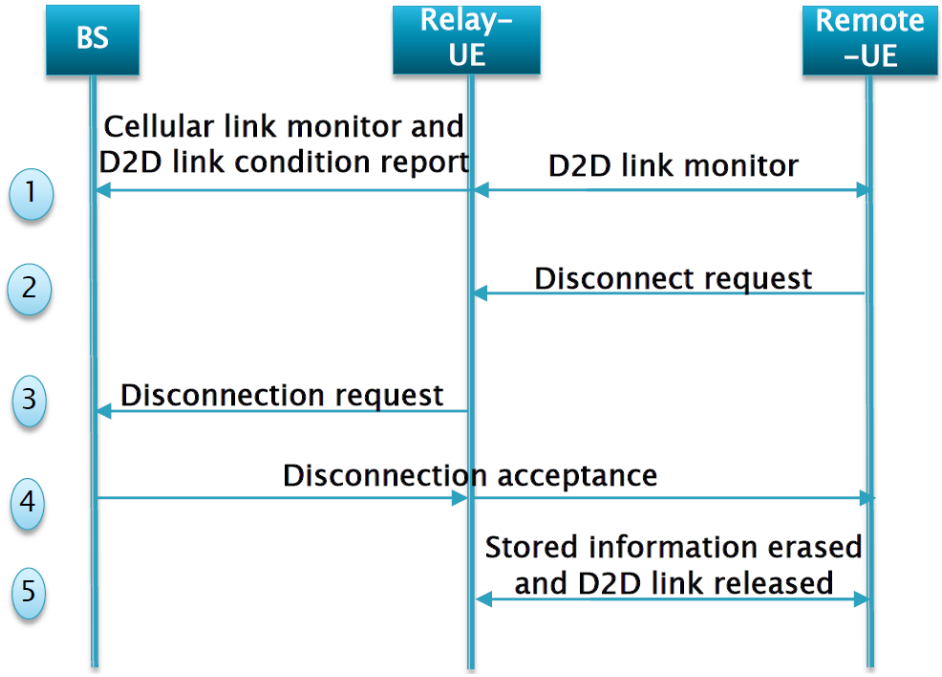


Figure A-9 D2D pair release procedure

1. BS and both D2D ends monitor the D2D condition between relay UE and remote UE, and also the cellular link between relay UE and BS.
2. If remote UE decides to release the sidelink, it transmits a disconnection request to the relay UE.
3. The relay UE will forward the disconnect request to BS.
4. After receiving the disconnect request, BS replies with a disconnection acceptance message.
5. D2D ends release the D2D link and erase their stored context information of this link.

After the release of a D2D pair, either a relay reselection procedure as stated before can be triggered by BS or the remote UE will be configured as a UE with a direct cellular connection in

uplink. The relay node previously used to serve the remote US should be excluded when a relay reselection procedure is performed.

As stated before, Figure A-9 shows the case where D2D releasing is originated by remote UE. However, if the releasing procedure is triggered by the relay UE, only step 3, 4 and 5 are required. Moreover, if the D2D pair releasing procedure is triggered by BS, step 1, 2 and 3 in Figure A-9 are skipped and BS can directly command two D2D ends to erase their stored context information and release the D2D pair.

A.4.2 Clustering methods

Geometrical clustering

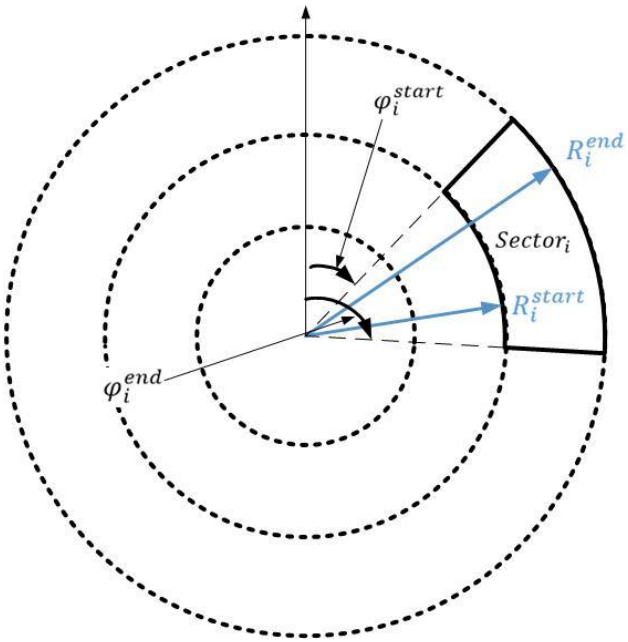


Figure A-10 Geometrical clustering

In this method, the coverage area of one BS is sectorized w.r.t. geometrical information, as shown in Figure A-10. The area covered by the *i*-th cluster can be represented by a radius of R_i and an angle of φ_j as

$$R_i^{start} < R_i \leq R_i^{end}$$
$$\varphi_i^{start} < \varphi_i \leq \varphi_i^{end}$$

where R_i^{start} and R_i^{end} represent the distances from the origin to inner and outer circles of the *i*-th cluster. Moreover, φ_i^{start} and φ_i^{end} represent the reference angles between which the *i*-th cluster covers. The number of clusters covered by one BS is a function of A_{sector}, which is the area of one cluster.

K-means clustering

K-means is one well-known clustering algorithm and its basic steps are listed below.

1. Initially, we select K points which are placed as far away as possible from each other and these K points are considered as centroids of the K clusters.
2. Take one another point and associate it to the cluster which has the shortest distance from its centroid to this point.
3. Calculate the mean point of the new formed cluster and select the nearest point to the mean point as the new centroid.
4. Repeat step 2 and 3 until all points are associated to a cluster.

It can be seen that the centroids of clusters are iteratively adjusted in this scheme.

Distance based clustering

This scheme is similar to the K-means clustering algorithm.

1. K points are randomly selected as centroids of K clusters.
2. Take one another point and associate it to the cluster which has the shortest distance from its centroid to this point.
3. Repeat step 2 until all points are associated to a cluster.

It can be noticed that, the centroids are selected randomly from the data set and these centroids are not updated during the operation of this algorithm. These are the two difference of this scheme compared with K-means clustering method.

Distance plus CSI based clustering

In this scheme, not only the location information of UEs are considered, but also the CSI from each UE to the BS. The difference compared with distance based clustering method is that, the K centroids are selected from the devices which have cellular SNR values higher than a pre-defined threshold.

A.4.3 Transmission mode selection

The battery life requirement of MTC devices can be up to 10 years. For UEs who cannot meet the battery life requirement, D2D communication is exploited. The equation below describes the condition of remote UEs whose battery life requirement cannot be met by cellular links:

$$\frac{BC_{(i,j)}}{EC_{(i,j)}} < BL_{threshold}$$

$BC_{(i,j)}$ denotes the battery capacity of user-*j* in cluster-*i* and $EC_{(i,j)}$ is the energy consumption of that user served by cellular link for a time unit of Δt. Moreover, users which cannot reach BS by cellular links can be assumed to have an infinite value of energy consumption for Δt. Thus, these users also fulfill the inequality and D2D communication is also applied to improve their availability. Last but not least, the requirement of battery life is denoted by $BL_{threshold}$.

If some UEs are classified as remote UEs in a cluster, BS checks whether certain UEs in the same cluster fulfill the following conditions for being relays:

$$\frac{BC_{(i,j)}}{EC_{(i,j)}} > BL_{threshold}$$
$$PL_{(i,j)}^{cellular} \geq PL_{threshold}$$

The first equation represents the condition that user-*j* in cluster-*i* can meet the battery life requirement by using cellular transmission. In other words, this user has enough battery capacity to serve as a relay for other remote UEs in cluster-*i*. In the second equation, $PL_{(i,j)}^{cellular}$ is the pathloss value of the cellular link and $PL_{threshold}$ is a predefined threshold value to check whether the channel condition of the cellular link is good enough. In this work, a value of 136dB is set as the threshold value. With a higher value for this parameter, UEs with better cellular link CSIs are considered as relay UEs and thus a higher spectral efficiency can be achieved for the relay links. However, higher value of $PL_{threshold}$ means less feasible relay UEs and there is a higher risk that no feasible relay UE exists in one cluster.

A.4.4 Evaluation methodology

In this section, models used in our simulator are stated with details. Please note that, only the difference compared with ITU-R performance evaluation document are given here. For other parameters not mentioned here, they are aligned with the ITU-R document [ITU09].

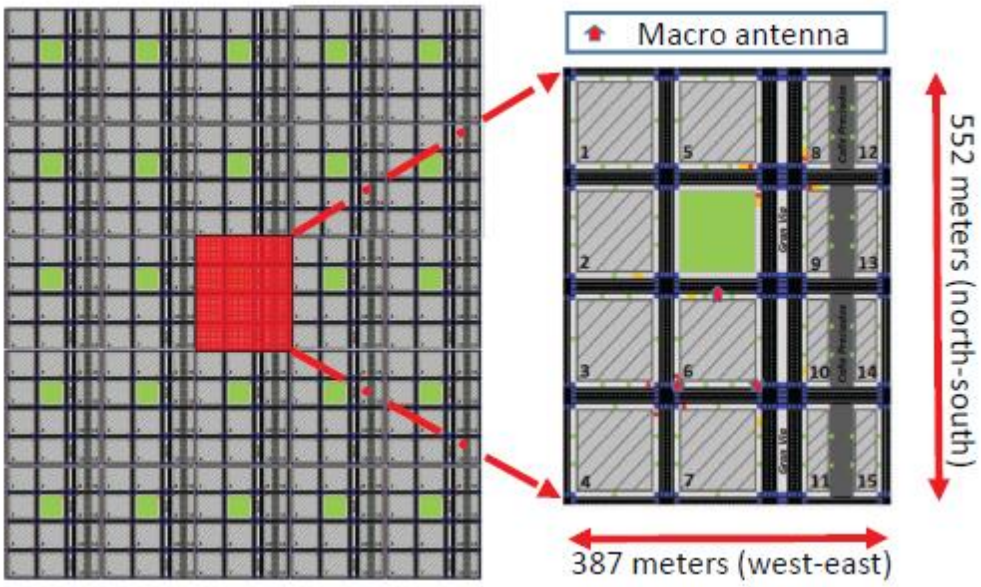


Figure A-11 Environment and deployment model

Environment model

In this work, system performance of the MTC service is investigated in dense urban environment as shown in Figure A-11 , where a Madrid grid model is applied [MET13-D61]. The proposed environment model aligns well with the reality to generate meaningful and precise results. In this model, an urban environment is depicted with 3D visualization where each grid composes of one park and 15 buildings of different dimensions. The dimension of one Madrid grid is 387 meters in west-east direction and 552 meters in north-south direction. In order to achieve a cell radius of 866 meters, multiple replicas of Madrid grid are generated in the system level simulator. Moreover, building heights in the Madrid grids are uniformly distributed between 8 and 15 floors with a height of 3.5 meters per floor.

Deployment Scenario

A single macro BS with a cell radius of 886 meters is deployed in the Madrid grids, in order to achieve an inter site distance (ISD) of 1732m as defined in 3GPP [3GPP17-38913]. The position of macro antennas is also plotted in Figure A-11 . For the macro station, it operates in three cell sectors with carrier frequency of 900 MHz and directional antennas are positioned with 120 degree difference from each other in horizontal plane.

User deployment and traffic model

In the coverage of the BS, 20 thousand sensor devices are randomly distributed inside buildings and are assumed to be static. An isotropic antenna is installed on each device at 1.5 meter height with a maximal transmission power of 23 dBm. Moreover, a report packet of 1000 bits with a periodicity of 288 reports per day (1 packet every 5 minutes) is exploited in this work as traffic model for sensor devices.

Channel model

A 3D channel model proposed by 3GPP [3GPP15-45820] is applied here, in which the penetration loss through building floors and walls is taken into account. To characterize channel in between two ends of one D2D communication, channel models proposed in 3GPP [3GPP13-R1132030] are applied. In [3GPP13-R1132030], channel characters are captured in three different scenarios for indoor UEs, i.e.,

- two D2D ends are on the same floor in the same building;
- two D2D ends are on the different floors in the same building;
- two D2D ends are in different buildings.

User device power consumption model

In order to evaluation the power consumption of MTC device, power consumption related parameters are listed in Table A-4. Please notice that, a new UE state called connected-inactive state is proposed to serve for 5G [MII16-D61], and thus the duration of control plane establishment is calculated based on this new technology.

Table A-4 Device power consumption parameters

parameter	description	value	time duration if applicable
P_{tx}	transmission power	45% PA efficiency plus 60mW/s for other circuitry	MCS and packet size related
P_{tx}	power to receive packets from remote UEs	100mW/s	MCS and packet size related
P_{paging}	power to receive paging command	100mW/s	10ms
P_{clock}	clock to obtain synchronization	100mW/s	10ms
P_{cp}	power consumption during the control plane establishment procedure	200mW/s	7.125ms
P_{sleep}	power consumption in sleeping mode	0.01mW/s	time of UE staying in

			sleeping mode
D_{rx}	UE wakes up to listen to paging	4 times/day	
C	battery capacity	5 Wh	

A.4.5 System performance

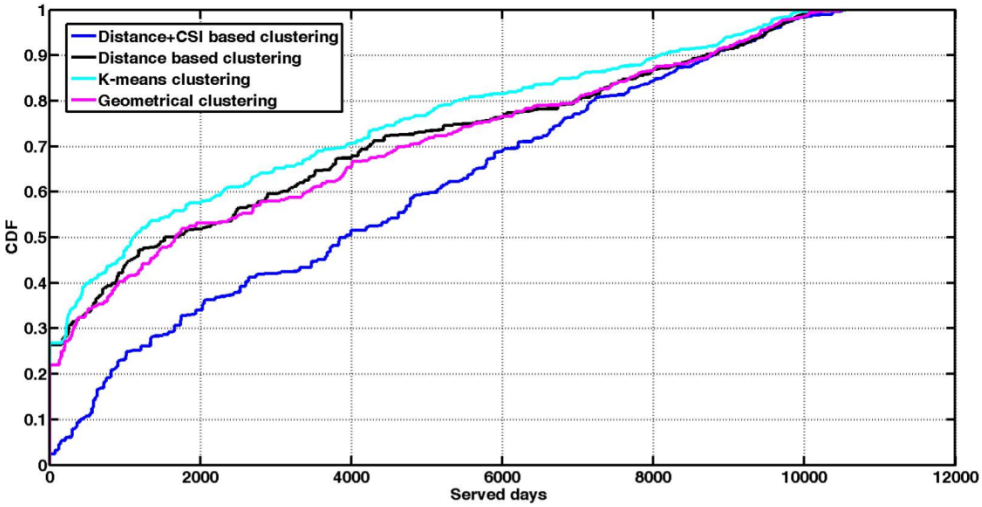


Figure A-12 Performance of the proposed schemes for users in outage of LTE

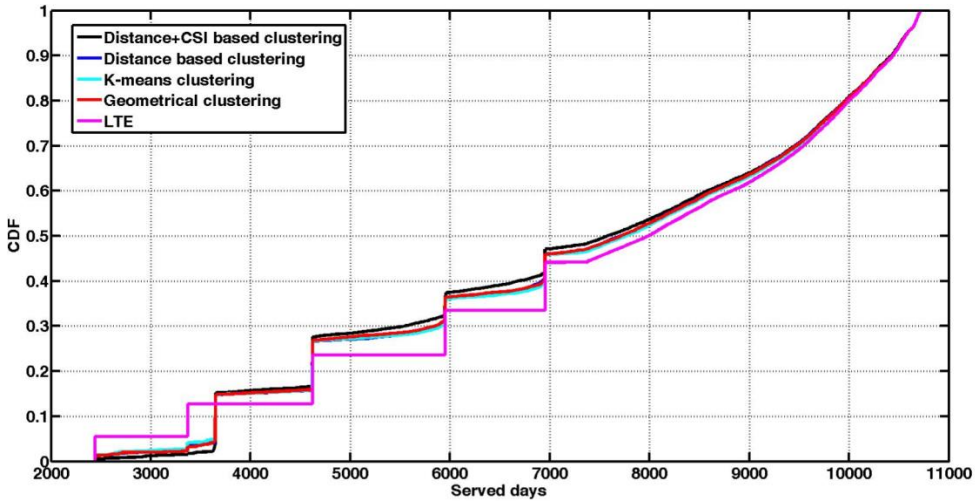


Figure A-13 Performance of the proposed schemes for users in coverage of LTE

A.5 Context-aware D2D underlay to improve system capacity

A.5.1 System model and RRM algorithm

In Figure A-14, D2D communication in reuse mode is shown.

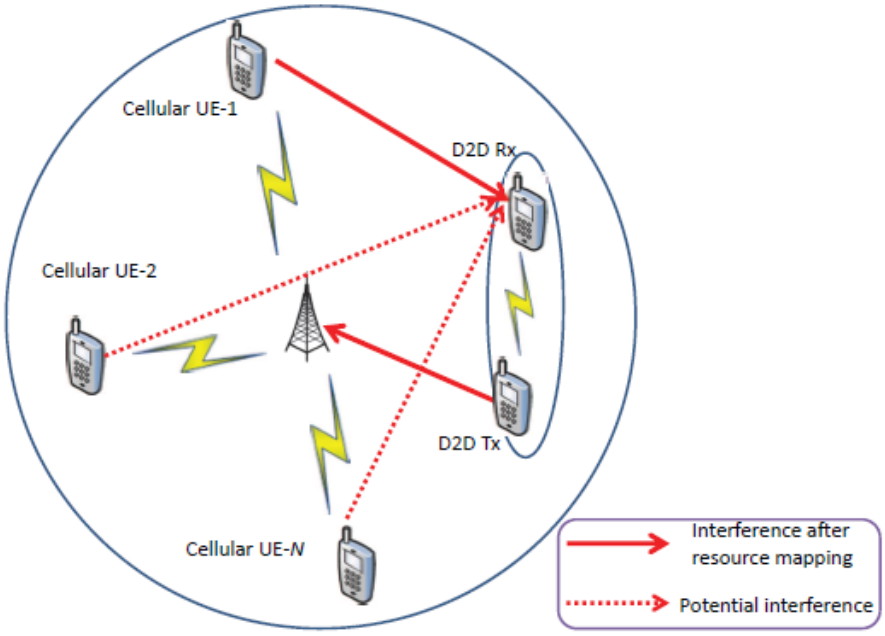


Figure A-14: Interference scenario in reuse mode

With total awareness of channel information, the BS can assign one cellular resource block to a D2D pair with certain consideration, i.e. to enable more D2D links or maximize the overall system capacity. In order to fulfil the QoS requirement of both cellular and D2D link, when one cellular uplink and one D2D link are assigned with the same time-frequency resource, the received signal quality should be taken into account by BS. Therefore, the *m*-th D2D pair can reuse resource block of *n*-th cellular UE, if and only if the SINR values of both the D2D link and cellular link are higher than the target values $SINR_{target}^{D2D}$ and $SINR_{target}^{cell}$, i.e.

$$SINR_{(m,n)}^{D2D} \geq SINR_{target}^{D2D}, \quad (1)$$

$$SINR_{(m,n)}^{cell} \geq SINR_{target}^{cell}. \quad (2)$$

The $SINR_{(m,n)}^{D2D}$ and $SINR_{(m,n)}^{cell}$ values are calculated as:

$$SINR_{(m,n)}^{D2D} = \frac{h_m^{D2D} P_T^{D2D}}{h_{(m,n)} P_T^{cell} + \sigma^2}, \quad (3)$$

$$SINR_{(m,n)}^{cell} = \frac{h_n^{cell} P_T^{cell}}{h_{(BS,m)} P_T^{D2D} + \sigma^2}. \quad (4)$$

P_T^{D2D} and P_T^{cell} denote the transmission powers of D2D and cellular transmitters, respectively. h_m^{D2D} represents the channel gain between transmitter and receiver of the *m*-th D2D pair, h_n^{cell} the channel gain between the *n*-th cellular UE and BS. $h_{(m,n)}$ represents the channel gain from the *n*-th cellular UE to *m*-th D2D Rx, $h_{(BS,m)}$ the channel gain from the *m*-th D2D Tx to BS. One channel gain value includes the effects from pathloss model, shadowing and antenna gain. σ^2 is used here to denote the thermal noise power per resource block, including influence of noise figure.

It can be seen from the equations above, CSI of overall cellular and D2D links are preferable if they are available at BS side, in order to support a smart RRM algorithm forD2D communications. A feasibility function can be constructed to show whether a D2D link and a cellular link can be assigned with the same resource while the SINR requirement are fulfilled for both the cellular and D2D link, as:

$$f_{(m,n)} = \begin{cases} 1, & \text{if both Eq. (1) and Eq. (2) are fulfilled;} \\ 0, & \text{else.} \end{cases} \quad (5)$$

In a single cell scenario with *M* cellular UEs and *N* D2D pairs, if D2D transmission is uni-directional, a total awareness of CSI means following information:

- *M* cellular link channel gain information (one link between uplink cellular UE and BS).
- *N* D2D link channel gain information (one link between D2D Tx and Rx).
- *N* channel gain information to represent the interference link for cellular UE (link between D2D Tx and BS).
- $M \times N$ channel gain information to represent the interference links for D2D links (link between uplink cellular UE and D2D Rx).

Thus, the channel gain information representing the interference links for each D2D link will be collected in a cumbersome manner. Meanwhile, in order to collect them, a signalling scheme with low efficiency will be brought up.

In order to achieve an efficient signalling scheme, we propose a scheme where UEs' positions are exploited by BS to decide whether one D2D link can re-use the resource of a cellular UE-n. In this scheme, a BS considers that interference from a cellular link to a D2D receiver is under control if following equation fulfils:

$$\frac{d_m}{d_{(m,n)}} \geq \gamma_{(SINR_{target}^{D2D}, d_m)}, \quad (6)$$

d_m is the distance between the two ends of the m-th D2D pair, and $d_{(m,n)}$ is the distance from the n-th cellular UE to the receiver of the m-th D2D pair. $\gamma_{(SINR_{target}^{D2D}, d_m)}$ is a threshold value which is a function of the target SINR value of the D2D link $SINR_{target}^{D2D}$, d_m and other context information if they have certain usage. With this way of approximating the channel information, BS will consider it is feasible for one D2D pair to reuse the resource of one cellular user, as

$$\tilde{f}_{(m,n)} = \begin{cases} 1, & \text{if both Eq. (2) and Eq. (6) are fulfilled;} \\ 0, & \text{else.} \end{cases} \quad (7)$$

In our work, we use a simple method to set the decision threshold for both cellular and D2D links to replace $SINR_{target}^{cell}$ and $\gamma_{(SINR_{target}^{D2D},d_m)}$ in Eq. (2) and Eq. (6) respectively, as

$$SINR_{(m,n)}^{cell} \geq SINR_n - \gamma_{cell}, \quad (8)$$

$$\frac{d_{(m,n)}}{d_m} \geq 1 \quad (9)$$

$SINR_n$ represents the SINR value experienced by the *n*-th cellular link when no D2D link reuses the same resource. γ_{cell} represents a deterioration offset allowed by RRM algorithm. In this work, the system is optimized w.r.t. the feasibility function defined in Eq. (7) to enable as many D2D links as possible. The optimization problem can be efficiently solved by the algorithm proposed in [JKK+14].

A.5.2 Evaluation methodology

Detailed information regarding the simulation models and assumptions are captured in [MET13-D61]. We highlight the most relevant models and parameters for our work in this section.

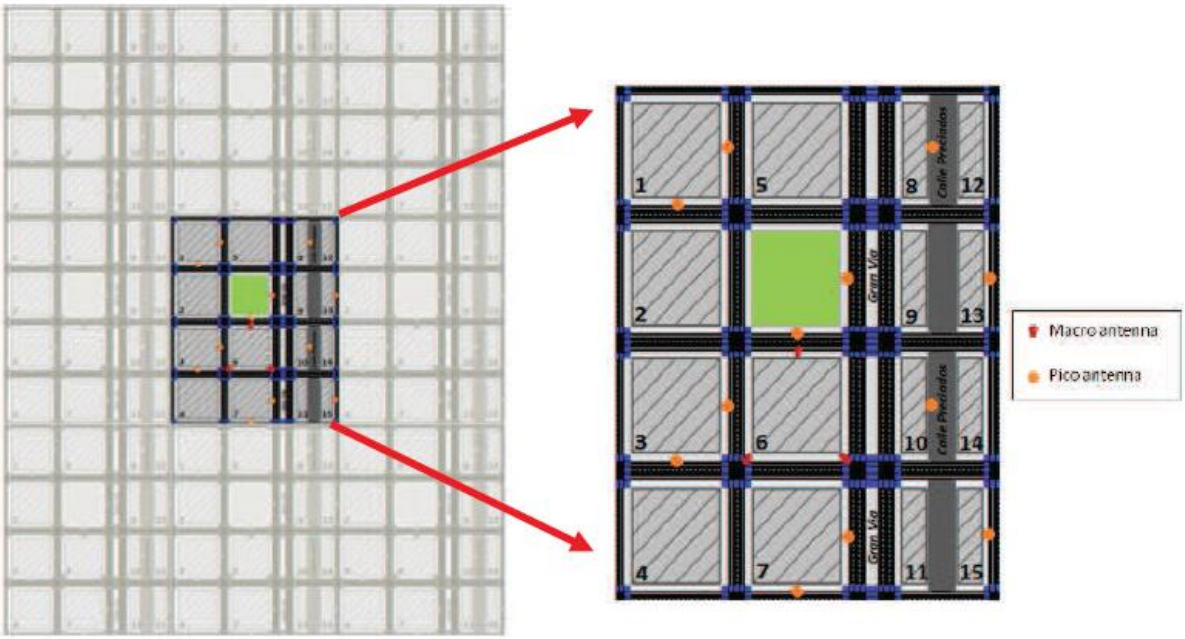


Figure A-15 Environment and deployment models

Environmental model

In this work, system performance of the proposed D2D communication is investigated in a dense urban environment. As shown in Figure A-15, a Madrid grid environmental model aligned with reality is used here [MET13-D61]. In this model, an urban environment is depicted with 3D visualization where each grid composes of 15 buildings with different dimension and one park. In order to avoid cell border effect, another eight replicas of Madrid grid are also placed but only the users located in the central Madrid grid are inspected to derive system performance. Two dimensions for one Madrid grid are 387 m (east-west) and 552 m (south-north). And building

heights are uniformly distributed between 8 and 15 floors with 3.5 m per floor. Detailed description of this model can be found in [MET13-D61]

Deployment model

In Madrid grid, we inspect on users which are distributed outdoor with a density of 1000 users per square kilometer. An isotropic antenna is installed on each user at 1.5 meter height with a maximal transmission power of 24 dBm. We do not inspect on a more advanced antenna configuration since it is not the main focus of this work. Besides, a heterogeneous network deployment is assumed with both macro and pico cells. For macro station, it operates in three sectors with a carrier frequency of 800 MHz and directional antennas are positioned with 120° difference from each other in horizontal plane. At the same time, pico cells operate at central carrier frequency of 2.6GHz with two cells per pico station where each pico cell point toward the main street. Bandwidths of 10 MHz and 40 MHz are respectively used as cellular uplink resource for macro and pico cells.

Traffic model

A full buffer traffic model is used for both D2D and cellular links. Moreover, users are randomly selected to generate either cellular or D2D link traffic.

Channel model

A 3D channel model proposed in [MET13-D61] is used in this work which combines both a simplified ray-based approach and a pure stochastic and geometric approach. This model is much simpler than ray tracing but still allows for a proper characterization of 3D environment in reality.

A.6 Cooperative D2D Communication

The formulation of the optimization problem and the simplified implementation for the proposed cooperative D2D method

Our main objective is to find $\mathrm{Q}_k^{(1)}[t]$ and $\mathrm{Q}_k^{(2)}[t]$, the covariance matrices of the designed pre-coders in phase 1 and phase 2, respectively, to maximize the long-term sum utility of the system, subject to long-term individual power and rate-gain constraints as well as a constraint on the interference at the BS, where the long-term individual power constraint is used to limit the energy consumption of each device; the long-term rate-gain constraint is used to ensure that each D2D pair achieves a larger rate through cooperation; and the interference constraint limits the interference that the D2D transmission may cause on the cellular system.

The long-term individual power constraint can be formulated as

$$\lim_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T}P_k[t]\le\bar{P}_k,\ \forall k, \tag{1}$$

where $P_k[t]$ is the power of DT k in time slot t, and $\bar{P}_k$ is the maximum average power of DT k. The long-term rate-gain constraint can be formulated as

$$\lim_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T}R_k[t]\geq\lim_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T}R_k^{(NC)}[t],\ \forall k, \tag{2}$$

where $R_k[t]$ is the rate of *k*-th D2D pair in time slot *t*, and $R_k^{(NC)}[t]$ is the achievable rate of the k-th D2D pair in time slot t. The constraint on the interference at the BS can be formulated as

$$\eta_1\mathrm{tr}(\mathbf{C}_{G,k}[t]\mathbf{Q}_k^{(1)}[t])+\eta_2\mathrm{tr}(\mathbf{C}_{G,b}[t]\mathbf{Q}_k^{(2)}[t])\leq\mathrm{IT}_k,\ \forall k,\forall t, \tag{3}$$

where $\mathbf{C}_{G,k}[t]=\mathbf{G}_{k,b}^H[t]\mathbf{G}_{k,b}[t]$, and $\mathbf{C}_{G,b}[t]=\mathbf{G}_b^H[t]\mathbf{G}_b[t]$. $\mathbf{G}_{k,b}[t]\in\mathbb{C}^{N_b\times N_t}$ is the channel between DT k and BS and $\mathbf{G}_b[t]=[\mathbf{G}_{1,b}[t],\ldots,\mathbf{G}_{K,b}]$, IT_k is the interference constraint value for *k*-th D2D pair, and $\mathrm{Q}_k^{(1)}[t]$ and $\mathrm{Q}_k^{(2)}[t]$ are the covariance matrix of the designed pre-coders in phase 1 and phase 2 for *k*-th D2D pair in time slot *t* , respectively. Hence, the long-term precoder design problem can be formulated as

$$\max_{\mathbf{Q}_k^{(1)}[t],\mathbf{Q}_k^{(2)}[t],\forall k,t}\quad\sum_{k=1}^{K}g\Big(\lim_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T}R_k[t]\Big)$$

subject to

$$\begin{aligned}
&\lim_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T}P_k[t]\leq\bar{P}_k,\ \forall k,\\
&\lim_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T}R_k[t]\geq\lim_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T}R_k^{(NC)}[t],\ \forall k,\\
&\eta_1\mathrm{tr}(\mathbf{C}_{G,k}[t]\mathbf{Q}_k^{(1)}[t])+\eta_2\mathrm{tr}(\mathbf{C}_{G,b}[t]\mathbf{Q}_k^{(2)}[t])\leq\mathrm{IT}_k,\ \forall k,\forall t,\\
&\mathbf{Q}_k^{(1)}[t]\succeq 0,\mathbf{Q}_k^{(2)}[t]\succeq 0,\qquad\forall k,\forall t.
\end{aligned}$$

where $g(\cdot)$ is utility function. Here, $g(\cdot)$ can be, for example, the identity function, which yields the average rate, or the logarithmic function, which leads to proportional fairness.

Since the problem is a long-term precoder design problem, the design complexity is high. To reduce the complexity of the design, it is desirable to propose per-time-slot (short-term) per-user designs (i.e., designs that depend only on the current time and can be computed in parallel for different users) that can achieve long-term performance guarantees. Hence, we decouple the problem into a sequence of sub-problems, each depending only on the parameters of a single time slot. The temporal evolution of the objective and constraints are recorded through the instantaneous queue states where we construct virtual queues, i.e., the data queue, the cooperative queue and the energy queue to record the temporal evolution of the long-term achievable rate, long-term rate-gain, and long-term power consumption. The resulting short-term sub-problems can also be computed in parallel for each D2D pair.

The overview of proposed cooperative D2D transmission technique is shown in the Figure A-16. In the figure, the problem can be further decoupled into parallel weighted-rate-minus-energy-penalty (WRMEP) maximization problems, each corresponding to a different D2D pair. And the

WRMEP maximization problem can be further solved by the conventional solutions of optimization problem.

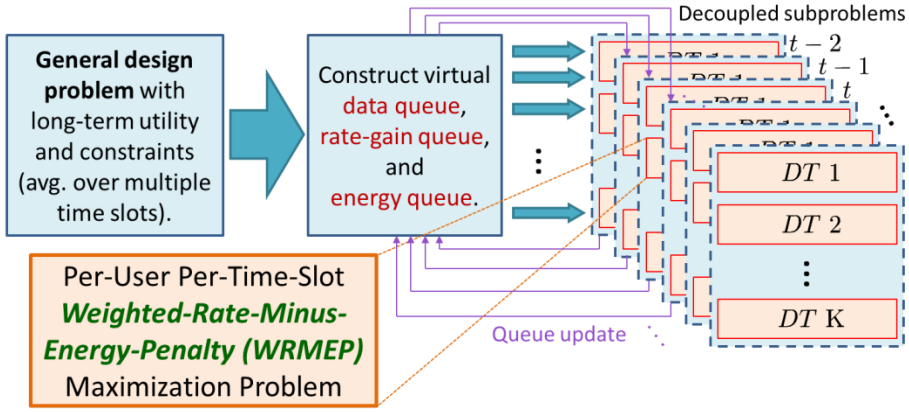


Figure A-16: Overview of proposed cooperative D2D transmission technique

Simulation parameters

As shown in Table A-5, both the circle and the ring are centered at (−150, −150). The number of antennas at all DTs, DRs, and CU is 3, and the number of antennas at the BS is 3K, where K is number of D2D pairs. The transmission is assumed to experience Rayleigh block fading and the path loss coefficient is chosen as 2. The results are averaged over 10 different user locations and iterations for each location. The time fraction is chosen to maximize the achievable rate under cooperation.

Table A-5 Simulation parameters

cell radius	$r_1 = 300$ m
center of D2D pairs	$(r_2, r_2) = (-150$ m, -150 m$)$
transmit antenna of DTs	3
receive antenna of DRs	3
transmit antenna of CU	3
number of D2D pair	$K = 3, 4, 5$
antenna of BS	$3K$
pathloss coefficient	2

A.7 Break before make handover in centralized deployments

A break before make (BBM) handover implies releasing the connection with the source cell before establishing a new one with the target cell. It is a viable approach for situations where the end-user service can tolerate an AS user plane interruption of couple of tens of milliseconds, and implies a low UE complexity. The BBM handover benefits from centralization in multiple ways, which are summarized in Figure A-17 below:

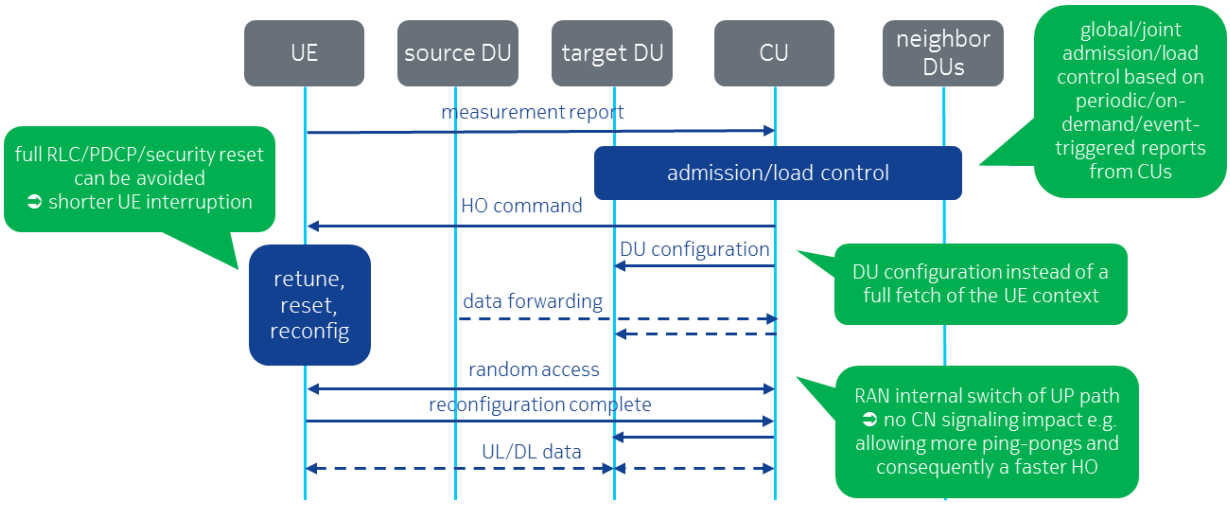


Figure A-17: Break before make handover in centralized RAN architecture

Centralization allows the admission control (AC) to acquire a global knowledge of the resource utilization of DUs, further allowing the load balancing (LB) and AC to be carried out in a joint manner. The resource status may be updated either periodically, upon request from CU, or upon a trigger configured by the CU, hence allowing the CUs to carry out a handover decision without explicitly querying the target DUs.

Centralization also allows optimizing the UE behavior during handover. The PDCP protocol and security can continue without a reset, hence decreasing the UE-induced service interruption. If the upper part of RLC is anchored to CU, also RLC re-establishment can be avoided. These benefits apply to both acknowledged and unacknowledged types of data bearer and the signaling bearer. The MAC and PHY layers are reset and the data is discarded as in LTE.

The UE context management benefits from the fact that part of the UE context resides in CU. In the case of a handover, the full context does not need to be transferred from the source to the target node, unlike in the distributed architecture, where a full UE context is fetched as part of the handover command. Instead, the UE context is modified as part of the handover procedure, configuring radio protocols in the target DU and releasing the ones located in the source DU.

Centralized architecture allows the user data path to be switched without CN involvement. This is unlike in the distributed architecture, where several signaling messages are exchanged between

RAN, MME, and sGW to switch the downlink data path. The reduced signaling load per handover allows utilizing a higher ping-pong rate, and consequently a handover with faster reaction time. It is particularly beneficial for high frequency bands which are shown in [3GPP17-R21700060] to experience a higher HO rate compared to low bands.

A.8 Fast activation of multi-connectivity

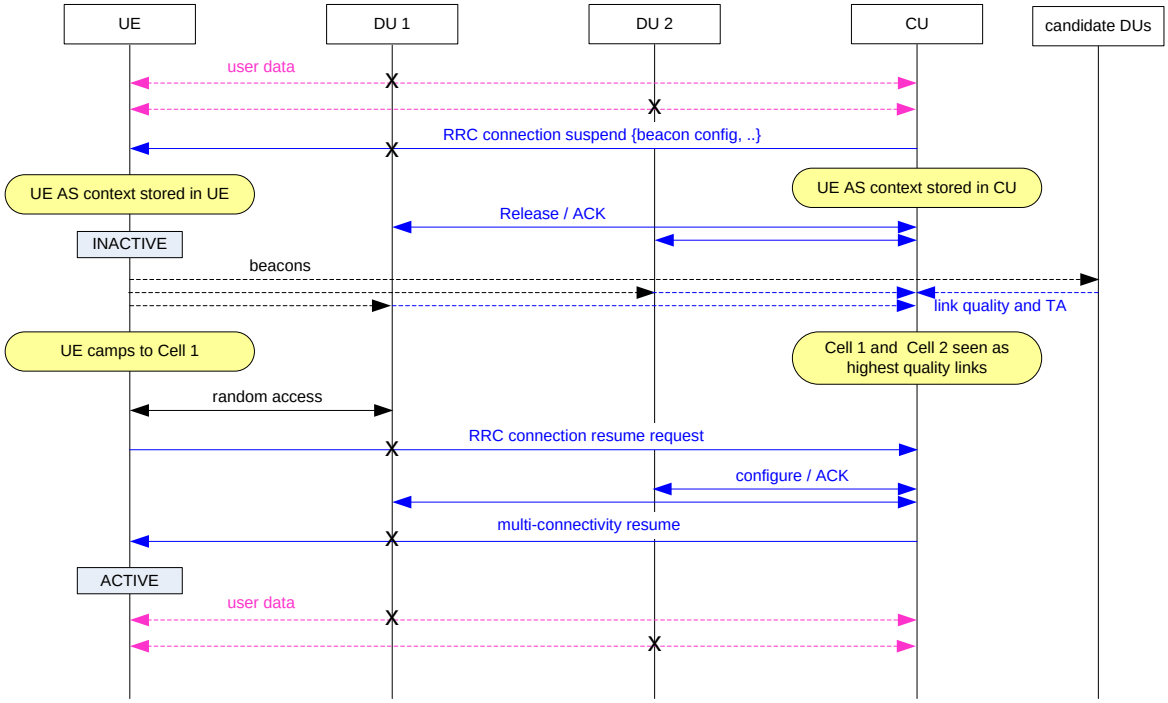


Figure A-18: Fast activation of multi-connectivity

A.9 Examples of CMP protocol use cases for 5G deployment scenarios

A.9.1 Self-Configuration

The term Self-Configuration encompasses the pre-operational actions that are performed before a network node starts to operate in the system (*self-establishment process*) as well as the operational actions that may be needed to modify the pre-operational parameterizations and configurations according to the dynamicity of the network (*self-adaptation process*).

In the 5G context, which will be characterized by a higher dynamicity and flexibility of the network nodes, significant technological and network deployment differences are expected. These can be summarized as follow:

- from one type of fixed installations where a radio base station is connected to the network using a cable, there will be a switch to a diverse set of different types of installation, where fixed, mobile or nomadic nodes can make the set-up of the backhauling via cable or via radio link. In this second case, the first connection to the network will have to take into account also the specificity of the mobile access, for example the available AIVs in a certain area, the authorization and authentication for the access through a public network, etc.
- in addition to the usual bands used for the mobile services, it is possible to assume the usage of the band in the millimeter waves (e.g. 30, 60 or 80 GHz), characterized by a very directive propagation and mainly in light-of-sight situations. In such a context, the instauration of a link could foresees also configuration actions related to, for example, the pointing of the antenna system before the effective connection to the network
- the 5G network will have to provide a number of variations in the functional architecture due to the separation of the control plane and the user plane both in terms of type of nodes (for example, the macro nodes dedicated mainly to the control plane and the micro/femto nodes to the user plane), both of used bandwidth used (with the traditional bands of the cellular system dedicated to the control plane functionalities and the millimeter waves to the user plane), both in terms of functional split (virtualized functionalities, centralized or decentralized for the two planes)
- the variability of the network topology will be increased by the possibility to activate/deactivate the access nodes for energy saving management

In the context just described, the parameter configuration of a network node presents, already in its initial phase, more degrees of freedom with respect to the actual situation in which every node has a predefined role and set of functionalities. In the operational phase, the 5G nodes may need a set of information (already obtained in the pre-operational phase) related to the following aspects:

- the functional procedures may be variable and dynamic according to the context and the time period of the day (for example, a node could operate as an access node for the user mobile devices during the day and being used for a machine type communication during the night)
- besides the traditional radio parameters for the mobility management of the users (for example for the hand-over between the cells of the control plane) handled through the traditional ANR SON functionality, the 5G network may need new cells clustering modalities for different scopes such as:
    - efficient spectrum allocation inside a group of nodes to reduce the interference

- the adoption of ESM algorithms (through the activation/deactivation of nodes) for energy saving purposes
- the support of moving nodes; in particular, inside a group of nomadic nodes, algorithms for the individuation of the best backhauling connection con be developed; similarly, the definition of appropriate node clusters could be targeted to identify an aggregator node for the user data flow for the uplink, or a distribution node for the user data flow for the downlink

The considerations reported above brings to assume a self-configuration procedure more complex and articulated as well as a set of information, exchanged between the network and the new node, more rich of context information.

The CMP protocol introduced in the previous section foresees that each network node has a context that contains the information that are required to manage in an efficient way a 5G network. In particular, such context information may include:

- Position (absolute o related to a cluster)
- Mobility (including speed): mobility type of the node (fixed, nomadic, mobile)
- Communication capabilities (AIV): number of possible AIVs, number of possible radio chains and their mapping with respect to the AIV in the appropriate transport planes; for each transport plane, the following additional information should be provided:
  - Connection status: active/non active
  - Throughput and latency performances, reliability of the connection
- Other capabilities: capabilities required for the execution of specific tasks such as, for example, elaboration capacity and storage
- Information about eventual belonging cluster; for each cluster the following additional information are needed:
  - Identification of the cluster (cluster-ID)
  - List of the other nodes of the cluster and the related received signal
  - Type/description of the cluster (for example, cluster of macro/micro nodes, fixed/mobile, purpose of the cluster such as energy saving, backhauling, aggregator, distributor, etc.)

From the logical/functional point of view, inside the CMP functional architecture is foreseen the implementation of a so called SCS (Self-Configuration Supervisor) functionality, having the scope to supervise the configuration by:

- Taking the decisions related to the network topology based on the information received from the nodes through the CMP protocol
- Maintaining updated the network topology status and the context of the nodes

The following sub-sections provide high level examples of the CMP protocol application.

Self-Establishment process

The initial configuration of a node can be defined through a procedure structured in different steps:

1) Connection establishment between the new node and the network (auto-connectivity) through a fixed or radio link. Once the connection is established, the SCS address has to be known or acquired.
2) The new node sends a Configuration Request to the SCS providing its position, mobility information, radio capabilities, other capabilities and the eventual belonging cluster.
3) The SCS sends a Configuration Set-up to the new node including the CMP-ID, software update/upgrade (for example to add or update new functionalities in the node, the supported AIVs, etc.), information on the topology of the connection (transport planes and related type such as backhauling, control plane, user plane, etc.), cluster information (including eventual new cluster identified by the SCS), radio parameters of the different AIVs (based on the AIV type, the operator policies, the specific transport plane, etc.).
4) The new node applies the configuration in terms of set-up of the backhauling protocol stacks, radio parameters and topology.

Self-Adaptation process

During the operational phase, the CMP protocol should support the reconfiguration of the pre-operational parameter. For example, the topology of a specific node, may be updated according to the network changes and can be defined through the following steps:

1) Monitoring of each transport plane (p1-pn) to collect the information related to the connection status and the throughput/latency performances.
2) Monitoring of the node in terms of connection status and mobility.
3) In case of changes that implies a modification of the network, the CMP protocol acts sending a Status Change message from the node to the SCS indicating, for example, the link status and/or topology changes.
4) According to the network status, the SCS takes the appropriate decision and sends a Configuration Update message to the interested node.
5) The node applies the updates according to what received form the SCS.

A.9.2 Self-Optimization

The automatic procedures for the optimization of the energy consumption of an access network by means of energy saving and energy efficiency mechanisms, as well as the automatic procedures for the optimization of the access network load through load balancing mechanisms, are part of the Self-Optimization process. The following sub-sections provide high level examples of the CMP protocol application to these mechanisms.

In relation to the energy efficiency mechanisms, it is worth to be noted that alternative solutions that act on a shorter time frame are investigated in Section 4.3 of the METIS II deliverable D5.2 [MII17-D52].

Energy Saving Management

From the logical/functional point of view, inside the CMP functional architecture is foreseen the implementation of a so called EES (Energy Efficiency Supervisor) functionality, having the scope of supervise the aspects related to the network energy saving and efficiency by:

- Maintaining updated the information (received through the CMP protocol) related to the energy efficiency (for example the consumptions) of the network and the nodes
- Managing the decisions (taken by appropriate mechanisms) related to the activation/deactivation of the nodes, traffic steering and network topology modifications on the basis of the information received from the nodes through the CMP protocol
- Monitoring and coordinating all the local and centralized energy saving mechanisms

Such functionality can be conceived as a separate entity that communicate with the SCS through specific messages of the CMP protocol. The EES could also superintend the specific energy saving mechanisms at a local level (cluster). In this case, the clusters could also operate in an autonomous manner, leaving to the EES only the task to activate/deactivate the energy saving features and collect the data related to the performances of the mechanisms. In this context, in case of different local energy saving mechanisms and/or centralized, the EES could act as orchestrator between such mechanisms, identifying eventual conflicts and the ones with the best performances. The CMP functional architecture foresees also the presence of an energy efficiency agent in each network node (EEagent).
In the management of the energy saving mechanisms, the EES (passing through the SCS) can acts on the network by activating/deactivating (or put in a sleep mode) the nodes, activating/deactivating the transport plans and modifying the network topology. It is worth to be noted that, in addition to the energy consumption information, the EES could ask to the EEagent in the nodes to monitor also other specific quantities typical of the energy saving mechanisms such as, for example, the CQI and the block probability at the node level. In particular, in the process of communication between the EES and the EEagent, the following phases can be defined:

1) Monitoring set-up: in this phase the EES retrieves through the CPM the node's characteristic (CMP-ID, computational/storage capabilities, EEagent capabilities in terms of quantities that is able to monitor and report) based on which configures the measurement to be performed.
2) Monitoring phase: in this phase, that generally coincides with the operational phase of the energy saving mechanism, the measurements useful for the algorithm are performed. The EEagent has the task to collect the measurements (eventually accordingly pre-elaborated). Reporting of the measurements can be done periodically, in occasion of a state change

internal to the node (appropriately monitored) or in response of a specific request coming from the ESS.

3) Monitoring deactivation: in this phase, when the energy saving mechanism is deactivated, the ESS close the monitoring activity by sending a specific command to the appropriate node(s) (identified through the CMP-ID)

The monitoring messages related to one or more energy saving mechanisms active in the network, may be received by the node in different times according to the decisions taken by the EES. This interacts with the SCS in order to obtain knowledge on the topology of the network and the status of the interested nodes. Based on such information, the EES can therefore modify the network topology, activate/deactivate nodes and connections through specific messages to the SCS in order to configure or modify the measurements to be performed. Such indications will be propagated to appropriate nodes through the Self-Adaptation process described in the previous sub-section.

Load Balancing

The exchange of specific information useful for the load balancing algorithms can be achieved through the CMP protocol. In particular, from the logical/functional point of view, the CMP functional architecture foresees the implementation of a LBS (Load Balancing Supervision) functionality that supervises the load balancing algorithms of the network by:

- Maintaining updated the information related to the load balancing (for example the network load and capacity) of the network resources and nodes
- Managing the decisions related to the activation/deactivation of the nodes, the traffic steering (of the access network and backhauling), the network topology modifications and the modifications of the cell's coverage (cell splitting, beam forming, radio parameters, etc.)
- Monitoring and coordinating all the local and distributed load balancing mechanisms

The LBS functionality can be conceived as a separate entity that communicates with the SCS through the CMP protocol. The LBS can also manage load balancing mechanisms at a cluster level. In case of local and distributed load balancing mechanisms, the LBS can act as an orchestrator between such mechanisms. The CMP functional architecture foresees also the presence of a resource load agent in each network node (RLagent).
It is possible to hypothesize that the network nodes would be characterized by specific "tags" (such as their capacity or capacity class) and a numerical value that indicates the load. More in detail, a metadata model that that completely characterizes a network node should contain at least two of the following three quantities:

- Load: load indicator in terms of a punctual value (for example PRBs usage in LTE) or quantitative (for example low, mid or high)

- Capacity: indicator in terms of numerical value of the available resources or in terms of capacity class (for example UL/DL relative capacity indicator)
- Percentage load/capacity: ratio between load and capacity in terms of percentage

It is then necessary to manage a database internal to the LBS in which the updated information related to the network nodes useful for the load balancing algorithm should be stored and maintained. Depending on the cases, such information could be:

- Provided by the manufacturers
- Retrieved from proper sensors in the network nodes interfaced with the RLagent and communicated to the LBS through the CMP protocol. Such sensors could be located in each node or in a specific subset of nodes
- Referred to an overall load (resource usage) of the network node, or distinguished between the various resources of the node such as radio access resources, transport resources (backhauling), base band resources, computational resources (useful for the edge computing) and storage capacity

In the case such information would be provided in an inhomogeneous manner e non directly comparable between them (for example in the case of specific radio resources used in different AIVs), transform tables between the quantities can be used in order to make the information usable by the load balancing algorithms inter-technology. More in general, the updates of the information in the database could be performed periodically or on an event basis. The addition of a new data could imply the update of the existing information, possibility maintaining an history to which the load balancing algorithm may refer for average analysis and/or forecasts.
Base on the previous considerations, the management of load balancing can be depicted in the three following main phases:

1) Monitoring set-up: in this phase the LBS retrieves through the CPM the node's characteristic (CMP-ID and the RLagent capabilities in terms of quantities that is able to monitor and report) based on which configures the measurement to be performed.
2) Monitoring phase: in this phase, that generally coincides with the operational phase of the load balancing mechanism, the measurements useful for the algorithm are performed and sent to the LBS through the CMP protocol.
3) Monitoring deactivation: in this phase, when the load balancing mechanism is deactivated, the LBS close the monitoring activity by sending a specific command to the appropriate node(s)

A.10 Context Aware Mobility

The scenario considered in is shown in left part of Figure A-19. The scenario has 25 crossroads and 7 landmarks (pinpoints). There are 6 coverage holes present in the simulation scenario at different roads shown as tunnels in the figure.

Right part of Figure A-19 shows an example of the application of this TeC where a vehicle is predicted to traverse a coverage hole.

In the simulations conducted for this TeC, there were 12 micro BSs with LTE-A technology (bandwidth of 10 MHz, 50 PRBs at 2 GHz carrier frequency). See more details in [KZS16].

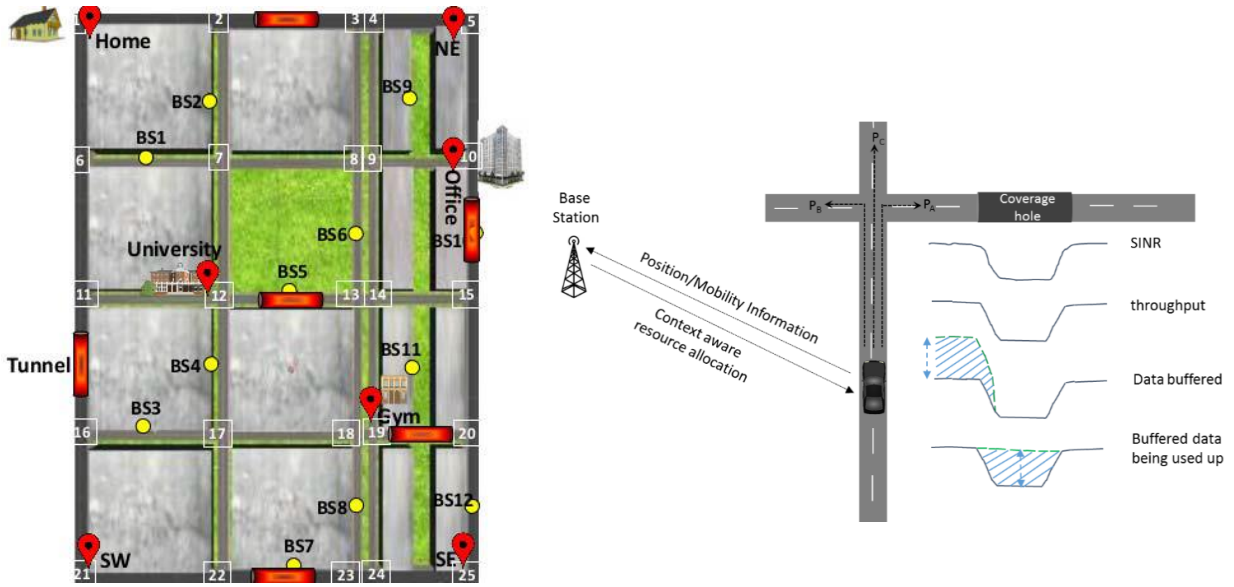


Figure A-19 Considered scenario with landmarks and coverage holes (left), and context aware radio RRM (right).