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

Deliverable D6.1
Draft Asynchronous Control Functions and Overall Control Plane Design

Version: v1.0

2016-06-30

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

Deliverable D6.1
Draft 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/D6.1
Document Title:	Draft Asynchronous Control Functions and Overall Control Plane Design
Version:	v1.0
Delivery Date:	2016-06-30
Editor(s):	Mårten Ericson, Icaro da Silva (Ericsson), Mikko Säily (Nokia Bell Labs), Panagiotis Spapis (Huawei), Shubhranshu Singh (ITRI)
Authors:	Mikko Säily, Jens Gebert, Tommi Jokela, Sofonias Hailu, (Nokia Bell Labs), Panagiotis Spapis, Alexandros Kaloxylos (Huawei), Icaro da Silva, Mårten Ericson (Ericsson), Shubhranshu Singh, Chorng-Ren Sheu (ITRI), Ji Lianghai, Nandish P. Kuruvatti (Technische Universitaet Kaiserslautern), Nico Bayer, Gerd Zimmermann (DTAG), Marco Mezzavilla, C. Nicolas Barati (New York University), Roy S Birdi, Jianjun Shen (Intel), Alessandro Trogolo (TI)
Keywords:	5G, Control Plane, RAN, Asynchronous functions
Status:	Final
Dissemination level:	Public

Revision History

Revision	Date	Description
1.0	2016-06-30	D6.1 Release v1.0

Executive Summary

This deliverable D6.1 presents the draft considerations on the asynchronous control functions and overall control plane design. It contains the preliminary (mid-point) view of the work package (WP) 6 of the 5G PPP METIS-II project.

This deliverable concludes that the overall Control plane (CP) architecture design for 5G must be able to fulfill a wide variety of requirements such as futureproof 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 AI 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 has to handle the new service requirement set by the industry, from use cases such as extreme MBB, URLL and mMTC. The report envisions that 5G should have a common CN/RAN interface (noted S1*) for both the new AIVs and the evolution of LTE-A, which enables a tighter interworking between the new AIVs and LTE-A evolution improving the mobility, robustness and resource usage.

In order 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 in order to cope with the sudden signal strength drop expected to occur in 5G. To support beam forming mobility efficiently, this report further proposes that new DL beam measurements should be complimented with UL

measurements. The beam forming mobility design should support a fast switching/tracking of the communication beam to combat rapid changes in link quality. Also, the design should be able to exploit the availability of multiple overlapping beams that can be used for the communication with a single UE. Further on, the beam mobility should have a minimum impact to the RRC layer. One solution to fulfill these requirements is the idea of cluster based mobility, which is a set of nodes that the UE can detect and which are prepared in advance for a fast re-routing of the signaling and user data. Also, context aware mobility solutions using data analytics are investigated. The benefits of the RRC Connected Inactive in terms of the mobility are reduced signaling towards the core network, power savings and lower latency.

The D2D functionalities in 5G are expected to be natively supported into the CP protocol stacks of novel AI or AIV, rather than as add-on functions, from devices as well as network perspectives. This requires several considerations and also imposes challenges from CP design perspective e.g. to support efficient & optimal resource management & channel access, unified addressing, cooperative communication, network offloading and inter-RAT/intra-RAT Mobility Management.

Contents

List of Abbreviations and Acronyms

3GPP	3rd Generation Partnership Project
4G	4th Generation
5G	5th Generation
5G-PPP	5th Generation Public-Private-Partnership
AI	Air Interface
AIV	Air Interface Variant
AP	Access Point
AS	Access Stratum
ASN.1	Abstract Syntax Notation
CA	Carrier Aggregation
CN	Core Network
CP	Control Plane
CQI	Channel Quality Indicator
C-RS	Cell specific Reference Signal
D2D	Device-to-Device
DC	Dual Connectivity
DL	Downlink
DM-RS	DeModulation Reference Signals
D-RAN	Dynamic RAN
DRX	Discontinuous Reception
DTX	Discontinuous Transmission
E2E	End-to-End
ECM	EPS Connection Management
eICIC	enhanced Inter-cell interference coordination
EMM	EPS Mobility Management
eNB	Evolved NodeB
EPC	Evolved Packet Core
EPS	Evolved Packet System
E-UTRAN	Enhanced UTRAN
FFS	For Further Study
GSM	Global System for Mobile Communications

HARQ	Hybrid Automatic Repeat Request
HO	Handover
HSPA	High Speed Packet Access
HW	Hardware
ICN	Information Centric Networking
ID	Identifier
IFOM	IP Flow Mobility
IoT	Internet of Things
L1	Layer 1
L2	Layer 2
LAA	License Assisted Access
LRU	Least Recently Used
LTE (-A)	Long Term Evolution (Advanced)
M2M	Machine-to-Machine
MAC	Medium Access Control
MMB	Massive Mobile Broadband
MC	Multi-Connectivity
MCG	Master Cell Group
MeNB	Master eNodeB
MIB	Master Information Block
MIMO	Multiple-Input Multiple-Output
MME	Mobility Management Entity
mMTC	Massive Machine-Type Communications
mmWave	millimeter Wave
MNO	Mobile Network Operator
MTC	Machine Type Communications
MU-MIMO	Multi-User MIMO
NAS	Non-Access Stratum
NF	Network Function
NFV	Network Function Virtualization
NGMN	Next Generation Mobile Networks
OAM	Operation and Maintenance
OTT	Over-The-Top

PCH	Paging Channel
PCI	Physical Cell ID
PDCCH	Physical DL Control Channel
PDCP	Packet Data Convergence Protocol
PHY	Physical layer
PLMN	Public Land Mobile Network
PSS	Primary Synchronization Signal
QoS	Quality of Service
RACH	Random Access CHannel
RAN	Radio access network
RAT	Radio Access Technology
RB	Resource Block
RLC	Radio Link Control
RRC	Radio resource control
RRM	Radio Resource Management
RS	Reference Signal
RSRP	Reference Signal Received Power
RSRQ	Reference Signal Received Quality
SCG	Secondary Cell Group
SDN	Software Defined Networking
SDT	Small Data Transmission
SeNB	Secondary eNB
S-GW	Serving Gateway
SIB	System Information Block
SON	Self-Organizing Networks
SSS	Secondary Synchronization Signal
TA	Tracking Area
TAL	Tracking Area List
TAU	Tracking Area Update
TCP	Transmission Control Protocol
TDD	Time Division Duplex
TTI	Transmit Time Interval
TTT	Time to Trigger
UDN	Ultra-dense Network
UE	User Equipment
UL	Uplink
uMTC	Ultra-reliable Machine-Type Communications
UMTS	Universal Mobile Telecommunication System
UP	User Plane
UTRAN	Universal Terrestrial RAN
V2X	Vehicle-to-Anything
VNF	Virtual Network Function
WG	Working Group
WP	Work Package
XaaS	Anything as a Service
xMBB	Extreme Mobile Broadband

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 eight billion [GSMA+16] 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 answer these control plane aspects of a 5G system.

1.1 Objective of the Report

This deliverable is a draft of the overall Control Plane (CP) design for the 5G RAN design concept proposed by the METIS-II project. Its primary objective is to report the initial 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. Despite the focus on the RRC design, the deliverable also contains considerations on the physical layer (PHY) protocol 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 LTE-A one. For example, the deliverable explains potential improvements and/or changes to the LTE-A design in order to support beam-based procedures (for mobility and initial access), the tight interworking between LTE-A and the new AIVs and energy efficient mechanisms. Note that the assumption in METIS-II is that the overall 5G AI will consist of multiple different AI variants (including LTE-A) – in METIS-II termed AIVs.

[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.

The technical solutions provided in this deliverable focus on the CP protocol design, which is typically standardized. Therefore, the deliverable may be potential input to ongoing standardization activities such those in 3GPP that have started in 2016.

1.2 Scope

The relation of this deliverable to other MEITIS-II deliverables is depicted in Figure 1-1. The reader is highly encouraged to refer to the stated other deliverables under [METIS].

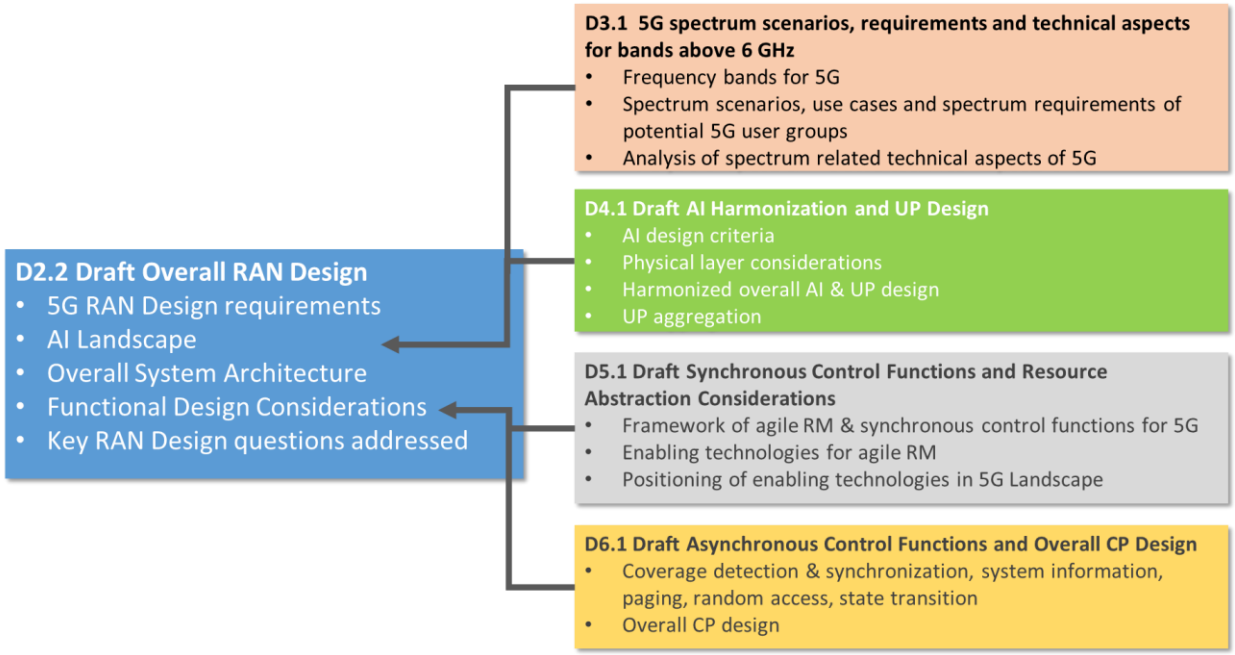


Figure 1-1 Relation of this deliverable to other deliverables.

As the successor of the METIS project, METIS-II inherits the METIS [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). [MET-II16-D51] describes the design of the synchronous functions while this deliverable is responsible for the initial design of asynchronous functions and the overall CP design. This deliverable also provides some initial conclusions concerning which functions will need to operate in which manner, i.e., which CP procedures related to mobility and system access are synchronous and which ones are asynchronous. The input from WP 5 has been recently provided in [MET-II16-D51].

The overall CP design will finally be integrated into the overall UP design in WP 2, responsible for the overall RAN design. Initial considerations of the UP and CP integration in the RAN design has been provided in [MET-II16-D22].

1.3 Structure of the Document

The document is structured as follows:

Chapter 2 presents the initial assumptions on the **Overall CP architecture**. The chapter contains the description of CP requirements such as future-proofness, the need to support high frequencies and beamforming, energy efficient mechanisms and tight interworking of the evolution of LTE-A with new AIVs, along with the standalone operation. The chapter also summarizes assumptions related to the overall architecture and presents the initial RRC design considerations, where most of the focus is on the CP solution for the UP aggregation of the evolution of LTE-A and the new AIVs.

Chapter 3 describes the new state handling for 5G, based on lessons learnt from LTE. A solution based on a new Connected (Inactive) state is proposed.

Chapter 4 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.

Chapter 5 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.

Chapter 6 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.

2 Overall Control Plane Architecture

2.1 Introduction

The 5G RAN architecture comprises the definition of the logical network elements, network interfaces for the overall air-interface. The air-interface is assumed to consist of multiple AIVs including the evolution of LTE-A [MET16-D22]. In METIS-II it is assumed that the overall 5G AI will consist of multiple different AI variants (including LTE-A). Examples of different AIVs could be different OFDM numerologies for different frequency ranges, but other variants are considered in the project and listed in [MET16-D22].

An essential part of the RAN architecture is also the CP architecture. This can be broken down into the following aspects:

- **Protocol architecture**, which includes the functional split between the CN and the RAN, the definition of logical network elements and network interfaces executing CP functions and the design of the full CP protocol stack, including the RRC protocol, often called the CP protocol.
- **CP procedures**, which includes for example the signaling exchanged between the network elements and the UEs (and its behavior associated to these messages) to support functionalities associated to initial access, mobility, scheduling, power control, interference management, etc.

This chapter mainly covers the requirements on the protocol architecture and CP aspects, while the remaining chapters focus on the CP procedures such as the ones associated to initial access and mobility. The chapter starts with an analysis of system level and performance requirements on the CP in Section 2.2, partially derived from [3GPP14-38913],[3GPP16-23799]. These requirements will later motivate the introduction of changes or enhancements to some of the LTE-A features. Then, the overall system considerations from [MET16-D22] are summarized, such as the CN design and CN/RAN split. Finally, the chapter describes the initial considerations on the RRC protocol design in Section 2.3 in order to fulfill the previously analyzed requirements.

2.2 CP Requirements

2.2.1 Future-proofness for Phased Standardization

Future-proofness is a system property that relates to the capability to introduce new features and services, while staying backward-compatible to the original design. This means that the CP design in initial 3GPP releases would need to be prepared for the introduction of new features and services that are difficult to predict what they are going to be since there could also be new requirements that are not yet being considered. To give one example, some level of future-proofness has been achieved in the LTE design, which can be acknowledged by a large amount of new features that were introduced after the initial release, e.g. enhanced Inter-Cell Interference Coordination (eICIC), Coordinated Multi-Point (CoMP), UE-specific Demodulation Reference

Signals (DM-RS), relaying, Machine-type Communication (MTC) enhancements (incl. Cat 1/0), Licensed Assisted-Access (LAA), Wi-Fi integration, Carrier Aggregation (CA), Dual Connectivity (DC) while still supporting multiplexing with legacy Release 8 UEs. In addition to these features, 3GPP has managed to introduce new services to the LTE AI, such as Massive MTC (mMTC) and currently Vehicle-to-Anything (V2X) communication is being standardized.

2.2.2 Operation in High Frequencies and Beamforming

To support the long-term traffic demands and efficiently enable the very wide transmission bandwidths needed for multi-Gb/s data rates, the range of operation might not only range to frequencies below 6.5 GHz (currently used for LTE) but also higher frequencies up to 30 GHz or above (e.g. 100 GHz). For that purpose, a wide range of spectrum with quite diverse range of characteristics, such as bandwidths and propagations conditions needs to be supported. In comparison to the current frequency bands allocated to LTE-A, most of the new bands have much more challenging propagation conditions, such as lower diffraction and higher outdoor/indoor penetration losses, which means that signals will have less ability to propagate around corners and penetrate walls. In addition, atmospheric/rain attenuation and human body blockage could also contribute to making the coverage of the AIVs in some of the new bands more challenging.

Beamforming, where multiple antenna elements are used to form narrow beams, is an efficient tool for improving both data rates and capacity. Its extensive use (in particular at the network side, is envisioned to be an essential part of the AIVs operating in higher frequencies in order to overcome the propagation challenges highlighted earlier. In that sense, the design of the CP procedures should take into account that for the higher frequencies narrow beamforming may be extensively used.

2.2.3 Key Performance Indicators for xMBB, URLL and mMTC

The support of the new 5G service categories, extreme Mobile Broadband (xMBB) and in particular Ultra Reliable Low Latency (URLL) as well as Massive Machine Type Communications (mMTC) imposes new stringent requirements in the CP design e.g. in terms of latency and reliability. To give some examples, as described in [3GPP14-38913]:

- UP latency down to 0.5 ms for both DL and UL should be supported for URLL.
- UP latency down to 4 ms for both DL and UL for xMBB.
- Reliability up to $1\text{-}10^{-5}$ within 1 ms for use cases such as eHealth surgical robots operating mainly in very deep indoor environment.
- User experienced data rate in the order of 300 Mbps in combination with the above mentioned reliability.

Another performance indicator relevant to the CP design is the CP latency which refers to the time to move from a battery efficient state (e.g., IDLE or INACTIVE) to start of continuous data transfer (e.g. ACTIVE). The target for CP latency should be 10 ms [3GPP14-38913].

KPIs associated to 5G use cases and especially mMTC are related to the massive number of accesses (which could be event driven). Specifically, the KPIs are:

- Efficient transmission of small data packets
- Efficient operation in bandwidth that is smaller than the total system bandwidth
- Connectivity for high density of devices, up to 1 million devices per km^2
- High service availability
- Long UE battery lifetime (up to 15 years)
- Minimum UP interruption time due to mobility procedures

The need to support these KPIs for these different services is reflected in the CP design as the need for high signaling reliability, fast and efficient initial access, fast activation of multi-connectivity, fast recovery from the radio link failures, minimum handover interruption time and fast and low signaling state transition (closely related to the CP latency).

2.2.4 Energy Efficiency

There has been a consensus among the different industry players involved in 5G R&D initiatives (such as [NGM15]) and research projects (such as [MET15-D84]) that reducing the energy consumption from radio networks is essential in 5G. Results from the EARTH project have shown that there is a potential for cutting energy consumption in current radio networks in the order of 50-75% [EAR-D64] indicating that some fundamental system design principles limit the maximum achievable energy efficiency. It has also been shown that most of the energy comes from the radio network does not come from generated traffic, but from the fact that radio nodes cannot efficiently move to sleep mode (even in low or no traffic situations) since there have to be some signals that are always on to make the CP work properly. A non-negligible part of the mobile operators' energy consumption comes from the physical layer processing and transmission even when there is no user data traffic in a given cell.

Therefore, the CP design should allow energy efficiency mechanisms without compromising the performance of certain procedures such as the ones associated to system access. One way to achieve this is by self-contained transmission of the system information and the use of a lean design of the reference symbols and broadcast information, see Section 4.3 for more information. Also note that the saving of energy using RRM algorithms is investigated in [MET-II16-D51].

2.2.5 Tight Interworking of LTE-A Evolution and the new AIVs

In order to enable mobile operators to leverage as much as possible on their previous investments in LTE, a tight interworking between LTE-A evolution and the new AIVs is assumed within METIS-II [MET16-D22]. Meanwhile, 3GPP has decided to evolve LTE-A which can be confirmed by the large number of enhancements proposed for Rel-14, to some extent associated to what has been identified as 5G use cases such as Vehicle-to-X (V2X) and Narrow Band Internet-of-Things (NB-IoT). Furthermore, it has been decided that the 3GPP submission to IMT 2020 (aka 5G) will include the new AIV(s), called in 3GPP the new radio, as well as the evolution of LTE-A. Hence,

it may be expected that the LTE-A evolution will also fulfil a significant number of the IMT 2020 requirements and will be widely deployed by the time that the new AIV(s) reaches the market.

The RAN architecture shall support tight interworking between the new AIV(s) and LTE-A evolution considering high performing inter-RAT mobility and aggregation of data flows via at least some sort of Dual Connectivity between LTE and the new AIV(s). This shall be supported for both collocated and non-collocated site deployments. In [MET-II16-D51] the performance gains of using PDCP aggregation are evaluated.

2.2.6 Support for Standalone Operation of new AIVs

The new AIV(s) should be able to operate in standalone deployments i.e. without any assistance from LTE-A evolution. That means that the new AIV(s) should have full CP and UP functionality such as initial access, mobility, multi-connectivity, support for different services, etc. and should be able to operate without interactions with the LTE-A evolution.

2.3 Initial Considerations on RRC Protocol

2.3.1 Assumptions on the Overall Architecture

In order to progress the CP design and design the RRC protocol, known as the CP protocol, it has been necessary to make some assumptions on the overall architecture and related interfaces. For the overall 5G Architecture, a CN/RAN split is assumed, aiming for an independent evolution of CN and RAN functionalities and allow multi-vendor deployments. In addition to that, a common CN and a common CN/RAN interface (noted S1*) for both the new AIVs and the evolution of LTE-A are envisioned. This enables a tighter interworking between the new AIVs and LTE-A evolution, improving the mobility, robustness and resource usage. Similar enhancements are also envisioned for the X2*, which jointly with S1* become interfaces addressing multiple AIVs. These assumptions are summarized in Figure 2-1.

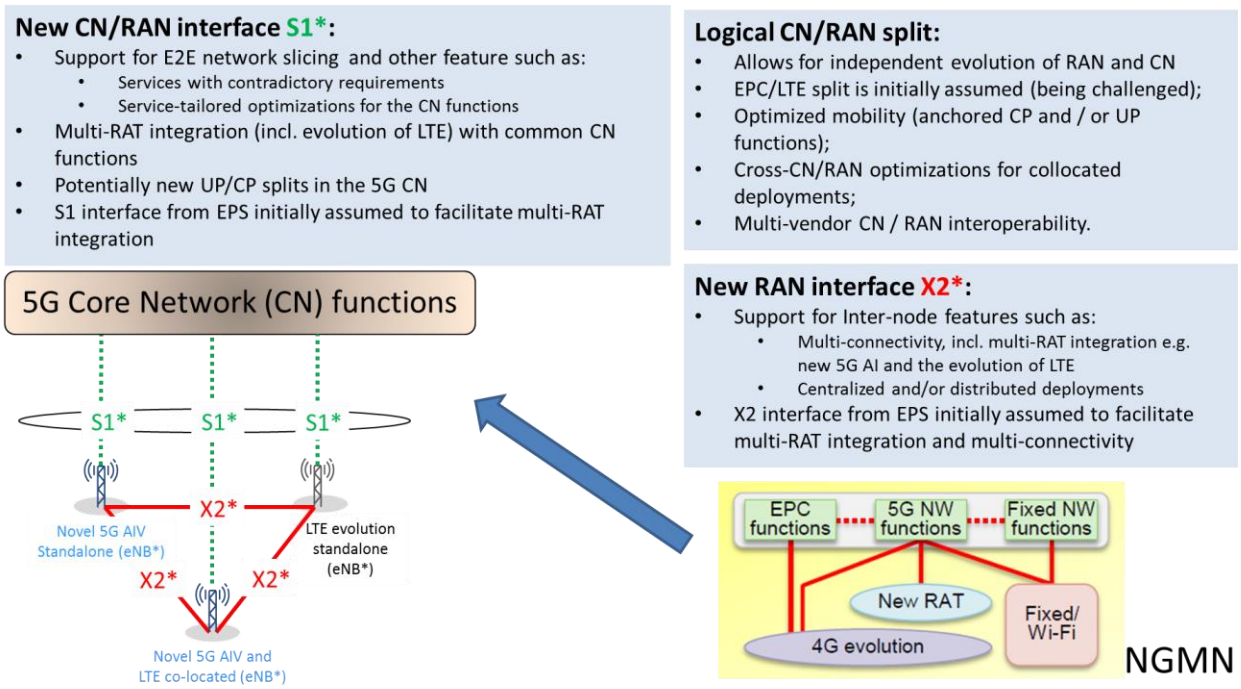


Figure 2-1: Baseline 5G RAN architecture assumed in METIS-II.

The fact that the proposed logical architecture has only inter-connected eNBs does not preclude deployments with further functional splits and/or centralized deployments. 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. More details can be found in [MET-II16-D22].

The RRC design in LTE-A has a simplified state model with an RRC CONNECTED and an RRC IDLE state. In RRC IDLE, the UE can be configured with a specific DRX, performs UE controlled mobility, monitors paging channel to detect incoming calls, system information change, neighboring cell measurements and cell (re-)selection, acquires system information, etc. In RRC CONNECTED, the main functions are the transfer of unicast data to/from UE, support for Carrier Aggregation and Dual Connectivity, support for network-controlled mobility, neighboring cell measurements and measurement reporting, etc.

As in LTE-A, the RRC for the overall 5G AI will continue to play an important role as the main CP protocol. Similar functions compared to LTE-A RRC are needed although the design of some of these functions might differ in order to cover new use cases, scenarios and fulfill some of the CP requirements listed in this chapter. Some of the main changes so far envisioned are the following:

- Support for a tight interworking between the LTE-A evolution and the new AIV(s), including mobility and at least Dual Connectivity;
- A new RRC state model natively relying on a lightweight connection;

- Support for a lean design where always on transmissions are minimized e.g. by new ways to distribute and encode System Information and/or new configuration mechanisms for reference signals;
- Support for beam-based procedures, both for initial access and mobility.

Standardization activities for 5G have started in 3GPP and perhaps one of the hottest topics is the support for the tight interworking between the LTE-A evolution and the new AIV(s), called the New Radio (NR) in [3GPP16-23799], more precisely the support for LTE/NR Dual Connectivity in RRC CONNECTED.

2.3.2 RRC Architecture for Dual Connectivity between LTE-A Evolution and new AIV(s)

Assuming the background of Dual Connectivity solution standardized from Release-12 in LTE-A, similar questions could be asked when it comes to the design of the RRC support for the tight interworking between the evolution of LTE-A and the new AIVs.

Coverage scenarios and the notion of an anchor AIV

The tight interworking between the LTE-A evolution and the new AIV(s) should support Dual Connectivity. The first thing to analyze is the coverage scenarios where Dual Connectivity applies. There will be scenarios where the UE has coverage only from one of the AIV(s) (e.g. in some indoor deployments in a coffee shop) and moves towards an area with overlapping coverage of the other. In order to support these two scenarios, where one of the AIVs could be the evolution of LTE-A the **UE should be able to connect via any new AIV and quickly establish dual connectivity with LTE-A evolution and vice versa**. Using the terminology borrowed from Release-12, in non-collocated deployments, either the eNB running LTE-A evolution or the eNB running any of the new AIVs should be able to act as some sort of Master eNB (MeNB). Therefore, the tight interworking solution will not imply the notion of an anchor AIV.

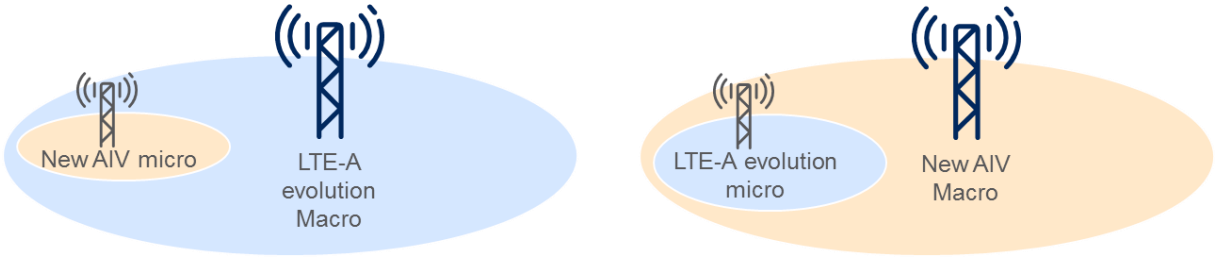


Figure 2-2: Scenarios where tight interworking between LTE-A evolution and the new 5G AIVs are applicable.

Assumptions on CN/RAN connectivity

One way to support this fast establishment from either LTE-A evolution or the new AIVs is to assume a single RAN/CN connectivity, i.e. single Non-Access Stratum (NAS), for LTE-A evolution and the new AIVs. This solution seems to be reasonable considering the assumption of a common CN and a common CN/RAN interface for LTE-A evolution and the new AIVs. Such a solution

makes it possible to have a common evolution for CN features of LTE-A and the new AIVs benefiting both AIVs at the same time avoiding separate specification work. In addition to that, a single NAS connection also simplifies the UE implementation, hence avoiding a dual protocol stack at the UE. When it comes to other procedures, it gives advantages when handling mobility and state transitions. In the case of mobility, a single handover procedure will be able to move the connections a UE has with each active AIV. In the case of state transitions, only one single CN state needs to be kept. UE, RAN and CN behavior due to such single state are greatly simplified and the risk of losing state synchronization is reduced. The main challenge is the need to align the CN/RAN signaling evolution of LTE and 5G.

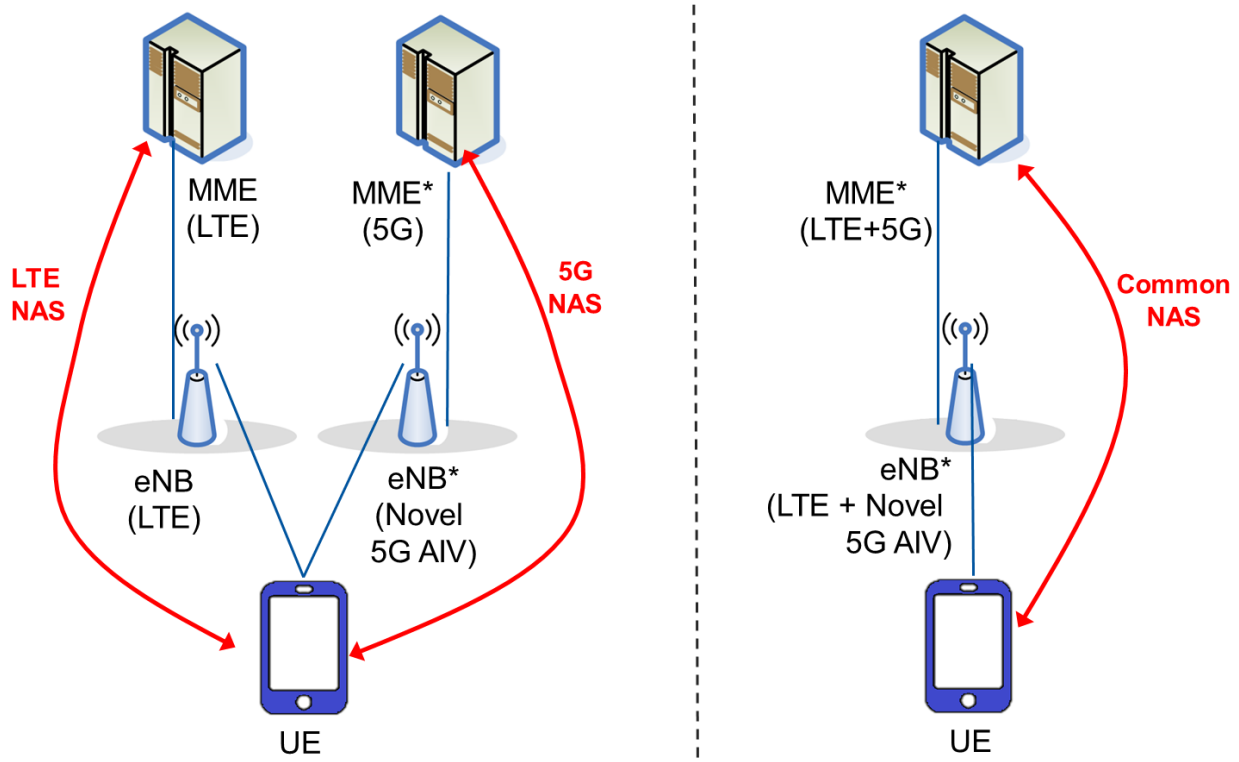


Figure 2-3: Dual NAS connectivity vs. single NAS connectivity for a UE using LTE and the novel 5G AIV simultaneously

Number of RRC connections and UE state

When it comes to the number of RRC Connections to the UE and the number of running RRC states, the project has investigated two alternatives. The first alternative is an evolution of the Release-12 solution where a UE maintains a **single RRC connection** with a single MeNB and has a single RRC state. In other words, the UE follows a single set of well-defined RRC procedures associated to a single RRC protocol such as establishment/modification/release of radio bearers, configuration of L1, L2 and L3 parameters and procedures. In case of Dual Connectivity setup, the procedure could closely resemble the signaling in LTE-A. By doing that, the RRC connection reconfiguration procedure for LTE-A and the new AIVs can be handled within a single-round of RRC message exchange. The solution also enables a lower complexity at the

UE side since the second AIV does not require the establishment of a new state machine. In addition to that, in this solution the MeNB decides the RRC reconfiguration of the L2 protocol and constructs the final RRC message conveyed to the UE.

In the second alternative, the UE maintains **dual RRC connections** with two eNBs and has two state machines running in parallel. One potential advantage of that approach is that the SeNB could configure its own resources directly by sending RRC reconfiguration to the UE, which in principle could reduce latency. However, some coordination should anyway occur between the different RRC entities from MeNB and SeNB anyway e.g., in order to sustain a single CN/RAN connection and coordinated state transitions between the two AIVs. Furthermore, even though the configuration of e.g., PHY layer has no impact on the CN/RAN interface or state handling, the UE capabilities may still need to be coordinated as they need to be shared between different AIVs. In summary, there can be similar amount or even more coordination with Dual RRC connection compared to single RRC which makes it a less likely solution to be adopted.

Number of RRC entities at the network side

It is obvious that in the case of a dual RRC connection there will be two RRC entities at the network side, as shown in Figure 2-4.

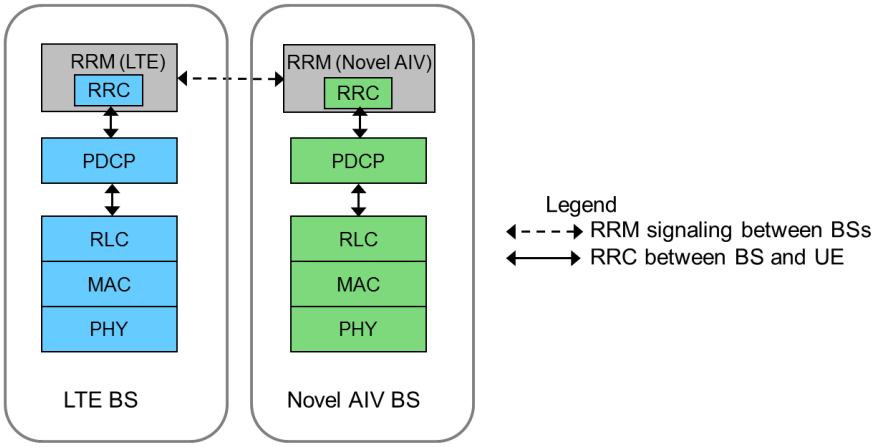


Figure 2-4: Dual RRC connection alternative with two RRC entities at the network

However, it is very important to clarify that a single RRC connection and a single UE state does not necessarily imply a single RRC entity at the network side but rather implies that the UE only sees and communicates with a single entity. In other words, either for single or dual RRC connection, as described previously, there could be one or two RRC entities at the network side generating ASN.1 associated to each AIV.

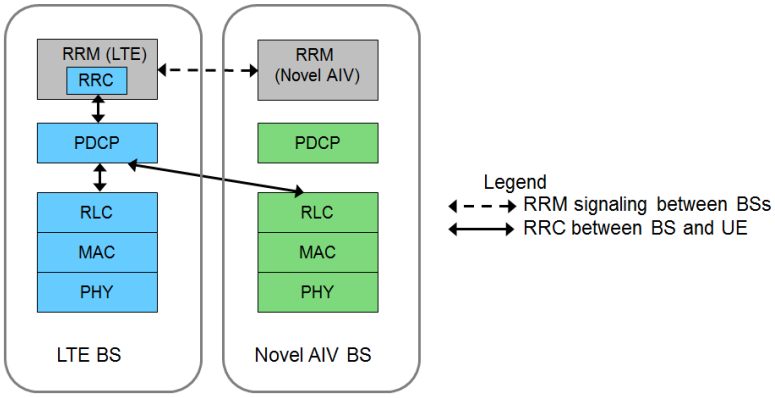


Figure 2-5: Single RRC connection alternative with two RRC entities at the network

Figure 2-5 shows the case where the UE sees only one RRC connection but where there are multiple RRC entities on network side. In such an evolved Release-12 solution, the RRC entity at the MeNB coordinates the CP actions and only a single entity of RRC generates the final RRC messages to the UE. In that case the SeNB generates the final configuration for its own resources as in Dual Connectivity i.e., there can be RRM related X2 coordination before the message is sent to the UE. One of the benefits of having two RRC entities at the network side is that it allows some level of separation between the RRC specifications of the LTE-A evolution and the new AIVs.

Transport of RRC messages in Dual Connectivity

In the case of a single RRC connection, the MeNB generates only one final RRC message. In the case of two RRC entities at the network, there could be different ways to transport that final message associated to two different AIVs.

One possible solution could be that information elements (IEs) associated to one AIV includes broadcasted or dedicated system information and security control information elements from the other, possibly generated in another eNB and coordinated over X2* (e.g., carrying radio resource control information elements), see Figure 2-6. These messages can be carried within RRC containers that need to be specified. In addition, the possibility should be studied to define these as transparent containers (especially from the SeNB to the MeNB) which would depend on how RRM functions are realized and what is the dependency of the RRC configuration in one and another node. As an example, IEs containing the configuration of new AIVs are carried in the Setup Response.

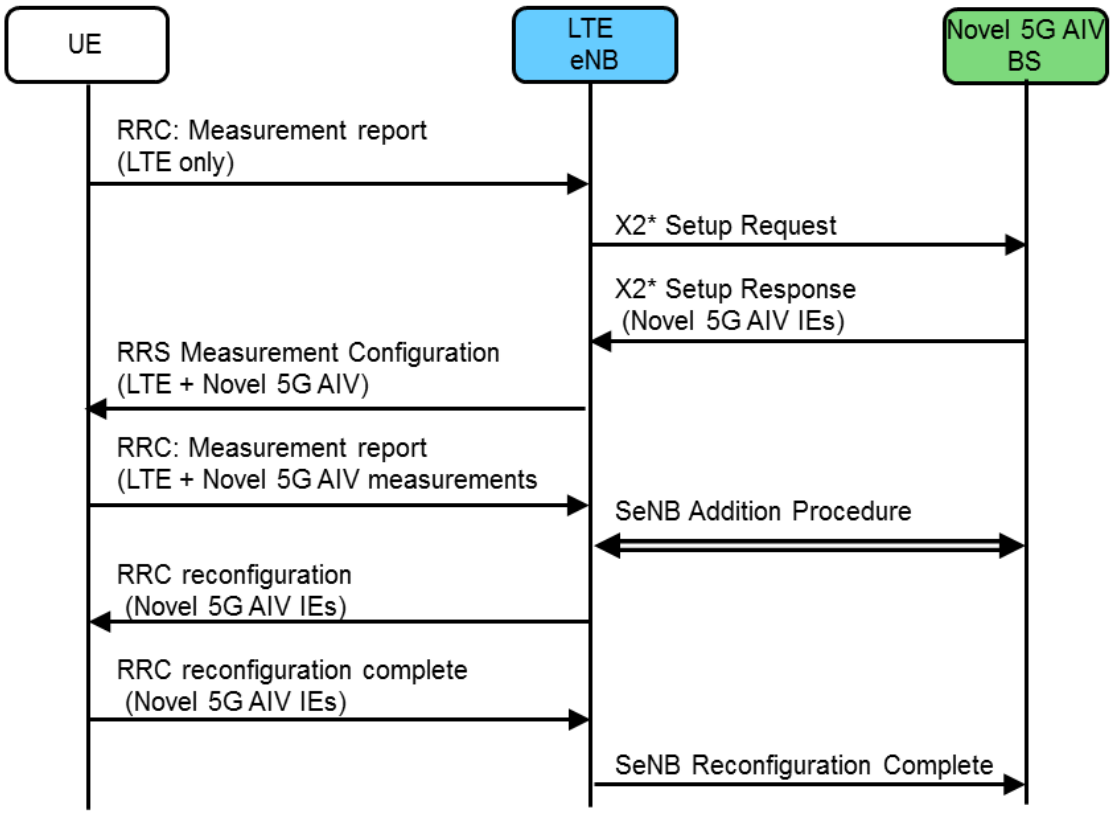


Figure 2-6: Single RRC connection for LTE and novel 5G AIV. Here, UE is first connected via LTE, then novel 5G AIV is added by using Dual Connectivity procedures. The reverse order is also possible.

Support for RRC Diversity and Fast CP switching

Considering that the new AIVs operating in higher frequencies will rely on beamforming where fast SINR drops may occur due to link blockage and higher penetration loss, mobility robustness becomes more critical than in LTE-A so that multiple routing algorithms for RRC messages become more attractive.

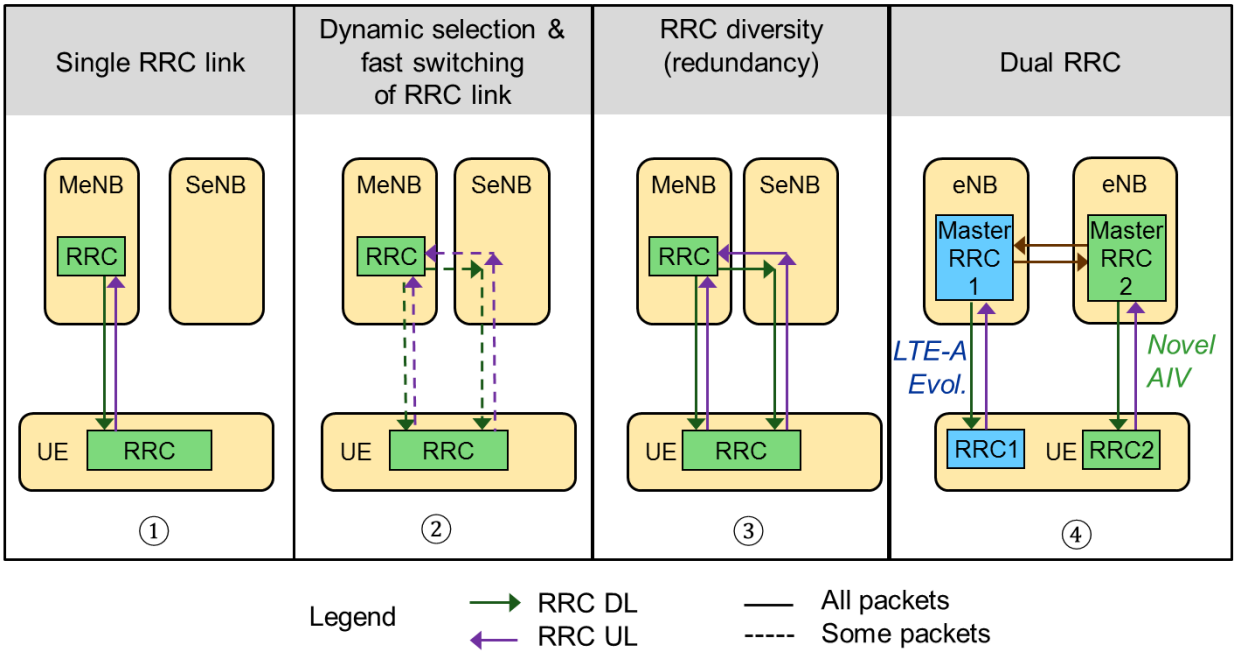


Figure 2-7: Different CP configurations with single RRC and dual RRC

As shown in Figure 2-7 the options ①, ② and ③ rely on a single RRC connection between the UE and the network (regardless of the number of RRM entities at the network) but different RRC routing configurations can be applied. In Option ①, the reference case, RRC messages associated to different base stations are transmitted over a single link which is the case supported by the current DC solution in LTE-A which relies on a MeNB RRC.

Option ② (RRC Dynamic Link Selection & Fast Switching) extends the previous option by routing the single RRC messages dynamically over any link, (e.g. based on radio conditions). The UE would be capable to switch very fast from one link to another (without requiring extensive connection setup signaling). The reliability might not be as high as in the RRC Diversity (where transmissions can occur simultaneously) and additional signaling is needed. On the other hand, one advantage is that the solution would work for UEs not necessarily having dual transceivers[3].

In option ③ (RRC diversity), RRC messages are duplicated and sent over each link. Such redundancy increases mobility robustness[4].

Option ④ shows the case where a UE has two or more RRC connections to the network. RRC Dynamic Link Selection & Fast Switching as well as RRC diversity can also be combined with this Dual RRC Option.

[3] However, devices using different AIVs/frequency bands (e.g. 2 GHz and 30 GHz) may need multiple transceiver chains anyway.
[4] At the expense of increased control plane traffic and possibly higher energy consumption.

3 State Handling

3.1 Introduction

The present chapter describes the envisioned RRC state model for the 5G RAN. The chapter starts with some background and the lessons learnt from previous systems in order to motivate the introduction of a new RRC state model natively relying on a *lightweight connection* for UE inactivity. In that context, a lightweight connection refers to the fact that when the UE starts an inactivity period it remains connected to network and keeps parts of the RRC context in order to later speed up the connection resume. The chapter also describes how this lightweight connection can be modeled as a new connected state, called herein *RRC Connected Inactive* [SMS+16], or as an extension of the RRC Idle state under standardization for the LTE-A evolution in Release-14 [3GPP16-RP160540].

Narrow-Band IOT (NB-IOT) is a technology by the 3GPP [3GPP15-RP-151621]. This technology is a narrowband radio technology specially designed for the Internet of Things. One special focus area of the methodology is a long battery life and support for a large number of devices. One of the proposed solutions in [3GPP16-TR23720] (Solution 5) is proposing optimization to UE Idle to Connected state transition, which re-uses information from the previous RRC connection for the subsequent RRC connection setup. The signaling optimization is introducing two new procedures called 'RRC Suspend' and 'RRC Resume' and introduces new states in the MME and in the UE called ECM_Suspended and RRC_Suspended.

The chapter is finalized with the considerations regarding the state models of the new AIVs w.r.t. the state model of the LTE-A evolution and potential features associated to the tight interworking.

The take-away of this section is to understand the rationale of the proposed state model 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 mobility and system access procedures of the new state model are configurable based on different aspects of use cases, device capability, access latency and security requirements, privacy, etc. The section also discusses some of the overall benefits of the novel state model and derives mobility-related signaling diagrams.

3.2 Background on State Handling

In LTE-A, at a RAN level there are currently two states: RRC Idle and RRC Connected [3GPP15-36300]. Each state covers a set of optimized procedures followed over the air interface for either battery savings or active data transmission. The signals and channels that the UE monitors, how to perform mobility, the way mobility should be performed, the way the UE should be contacted by the network and the way the UE contacts the network characterize these procedures. One of the most critical challenges to be taken into account when states are being designed is how to enable a fast transmission of the first packet either when the UE is returning from the RRC Idle

or within the state optimized for data transmissions. This tradeoff between power savings and system access latency is called herein UE sleeping problem. [SMS+16]. This problem is explained in the following, as a background to motivate the design of the RRC states for the 5G RAN.

3.2.1 Control Plane Latency

In order to enable the UE to save battery it is very important that the UE can switch off its receiver and/or transmitter (or parts of them) when there is no data to be transmitted or expected for that UE. This makes it possible to achieve significantly longer standby times in the UE compared to "talk" (or active) time. At the same time the UE should be reachable by the network (e.g. via paging) and, if it wants to transmit data, it should be able to quickly access the system and transmit its first packet. UE sleeping is currently supported in LTE-A by configuring the UE with DRX [3GPP11-36321]. A DRX cycle consists of "on periods" during which the UE monitors Downlink (DL) channels and physical signals (so that it can be reached by the network via paging among other actions) and "sleeping periods" when the UE can switch its receivers off. When the UE is utilizing DRX it probably also uses Discontinuous Transmissions (DTX), however, this is not specified in the standards. In LTE-A, DRX supports both "RRC Connected" and "RRC Idle" states. RRC Idle state is often seen as the primary UE sleeping state in LTE-A, where the network procedures and UE behavior have been optimized for power savings. In RRC Idle state the mobility is UE-based and does not require the transmission of measurement reports and the network contacts the UE via paging.

A quick system access and fast transmission of the first packet could be achieved by maintaining all UEs in RRC Connected with DRX. However, the RRC Connected procedures are optimized for data transmission rather than low battery consumption so this becomes inefficient for slightly longer inactivity periods. Instead, it becomes more efficient to release the UE to "RRC Idle" after a configurable inactivity timer is expired. The consequence of relying on releasing the UE to RRC Idle is that a quick transmission of first packet requires an optimized state transition from RRC Idle to RRC Connected. The importance of this performance indicator is reflected by the definition of a KPI called CP latency, which is the time it takes to the UE to perform a state transition from a sleeping state to an active state optimized for data transmission.

3.2.2 Lessons Learned from Existing Technologies

In addition to the UE sleeping problem, the RRC states design for 5G should take into account the lessons learnt from the pros and cons of state handling design for the existing technologies and take into account the 5G use cases and their requirements. Comparison of RRC states in existing technologies has been discussed in [3GPP16-R2-163441]. One of the observations was that the 5G design should keep to a very low number of RRC states to avoid extra complexity from states that have rather similar functionality. Instead of multiple states, it is proposed to have fewer states that can be configured to address multiple contradicting requirements. 5G scenarios and requirements [NGM15] indicate that the state design needs to fulfill the CP latency requirement and minimize the signaling overhead for frequent state transitions from low activity to full active state considering the expected trend in 5G to support frequent small packet data

transmission and reception. 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.

RRC states in HSPA

There are five RRC states in HSPA, which are a combination of system access and mobility related attributes such as control of mobility procedures, UE monitoring the dedicated physical channels, location update procedure and if the uplink data transmission is allowed or not in current state. In HSPA, the RAN Context Information is stored in the CELL_PCH and URA_PCH low activity states of connected state, and Context is released when UE goes to idle state.

From the HSPA state machine, we can observe that CELL_PCH and URA_PCH have almost identical characteristics [3GPP15-25303], except that the location of a UE is tracked at the cell level in the former case, while tracked on the URA level in the latter. Hence, HSPA seems to have multiple RRC states with overlapping features, which increase the system design effort and adds complexity to the implementation. With fewer RRC states in 5G the overlapping RRC state attributes could be avoided and thus minimizing the number of RRC states.

The RRC connected low activity states in HSPA also have benefits for fast system access with diverse applications and requirements. For example, the CELL_PCH state is a RRC Connected state but the UE is not having allocation for a dedicated physical channel and uplink transmission is not possible. The UE is using DRX for power saving and monitoring the Paging Indication Channel (PICH) for any incoming data from network. The benefit of CELL_PCH state is that the UE is connected to network during the low activity periods and position of the UE is known by UTRAN on a cell level. In URA_PCH state, the location of the UE is known on UTRAN Registration area level. These states enable faster system access without state transition from idle to connected mode.

RRC states in LTE

There are two RRC states in LTE, RRC_CONNECTED and RRC_IDLE. A UE is in RRC_CONNECTED state when a RRC connection has been established and otherwise the UE is in RRC_IDLE state. It has been observed in LTE networks that the inactivity timers releasing UE to Idle state are typically configured to be quite short (down to 10-60 seconds) which leads to a high amount of transitions from Idle to Connected. This state transition is quite costly in terms of signaling considering that the majority of the RRC connections in LTE transfer less than 1 Kbyte of data and then move back to RRC Idle [3GPP11-36822], [3GPP11-22801].

Figure 3-1 presents the states and possible state combinations in LTE-A for both NAS and AS according to the UE status.

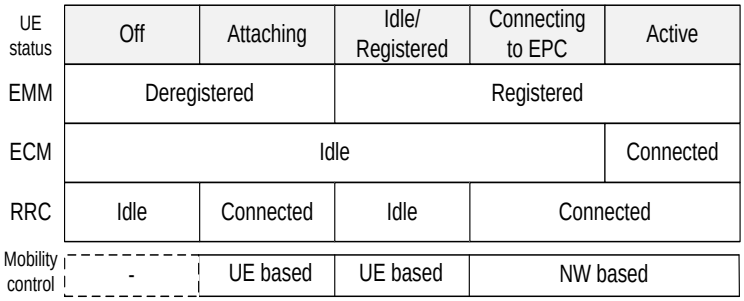


Figure 3-1: LTE-A Connectivity and RRC states

In the CN state (called EPS Connection Management (ECM) Idle / RRC Idle) only CN context is stored when the UE is sleeping. UE and network discards the RAN context information when moving to this state. The UE location is known at the network only on Tracking Area (TA) level and UE may move within the cells belonging to a TA List without informing the network. Nevertheless, the UE camps in the best cell via cell reselection procedure based on the configuration provided by the network [3GPP15-36300].

In the RAN connected state (ECM Connected / RRC Connected) with UE configured DRX the UE location is known on a cell level and mobility is fully network controlled (via handovers). In that state the RAN context is present.

The signaling diagram of Figure 3-2 shows the state transition from RRC Idle to RRC Connected, required for the UE to transmit/receive UP data. This overhead may also introduce significant delays. In the best case, without taking into account the processing delays at the UE, network side and signaling within the CN (e.g. towards the Mobility Management Entity and Serving Gateway), this overhead introduces a delay, measured in terms of Random Access (RA) and Round-Trip-Times (RTT) of:

Transition time > RA delay + 4 x RTT (radio) + RTT (S1 setup).

In LTE Release 8 this transition time should be lower than 100ms and, in Release 10, lower than 50ms [3GPP14-25912], so that even lower values in the order of 10ms should be expected in 5G at least for some services [3GPP14-36913]. Diverse service requirements are expected to exist in 5G networks especially with the trends towards massive IoT [3GPP15-45820], where it is expected a large number of devices, each generating a small amount of data.

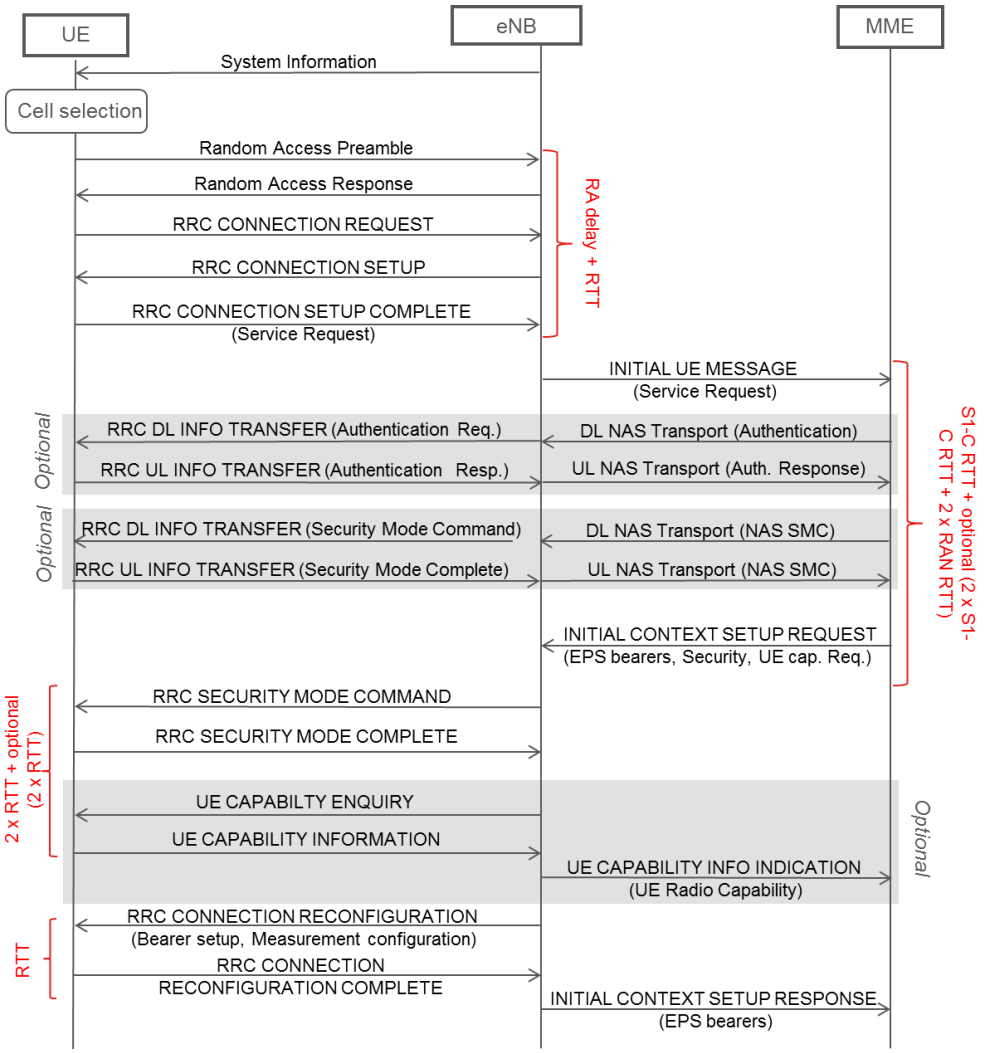


Figure 3-2: State transition in LTE-A (Release 10).

As can be observed from LTE RRC procedures, most of the radio signaling comes from random access and RRC connection setup procedures while the inter-node signaling, between the RAN and the CN comes from the whole setup of the S1 connection. The inactivity timers are typically configured to short values, in the order of 10 to 20 seconds. Therefore, for bursty traffic or mMTC devices the state transitions would occur in the same eNB meaning there is a potential to reduce the signaling if UE could stay in Connected for a longer period with low power dissipation. However, when the LTE RRC is examined it can be noticed that the RAN UE context is released during the idle state. Further, the UE inactivity timers result a high number of RRC state transitions, RAN context creation and release procedures. This transition from Idle to Connected state can have detrimental effects in terms of signaling overhead, as the majority of data connections in LTE transfer less than 1Kbyte of data before reverting back to the Idle state and consequently releasing the RAN UE context.

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 RRC-CONNECTED should be instantaneous from the end user perspective [NGM15], 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].

3.3 RRC Connected State

3.3.1 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 will have to support a growing set of static and nomadic users and devices. That is, mobility in 5G use cases indicates that in some services the inter RAT mobility functions can be disabled. Additionally, 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 local scale.

3.3.2 Characteristics of Connected State

In connected state the RRC connection has been established between the UE and the Network and the logical dedicated unicast resources are available for the transfer of control plane signaling or user plane data in uplink or downlink. 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 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 all the time to decode the incoming downlink data, as the data in the downlink may arrive at any time based on scheduling from network. The UE will be monitoring dedicated channels to check if there is downlink data available.

3.4 New RRC Connected Inactive State

In order to achieve a seamless UE state transition in 5G systems, a connectivity solution where the UE is kept 'always ON' from the Core Network (CN) perspective is considered. As can be seen in Figure 3-3 , once the UE is registered, 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. Some of the key features of the connectivity solution are as follows [3GPP16-S2161323]:

- The UE is 'always ON' from the CN network perspective
- The mobility and reachability functionality for UEs in RRC Connected Inactive state are handled by the RAN, leading to a new RAN-CN functionality split for inactive UEs
- From the RAN perspective, the RRC connection of a UE can be suspended keeping some of the UE context information, e.g. AS security context, in the UE and RAN
- The behavior of a UE in RRC Connected Inactive state can be configured based on the requirement of services that are provided to the UE.
- A light-weight signaling procedure is used to suspend and resume UE RRC connected state.

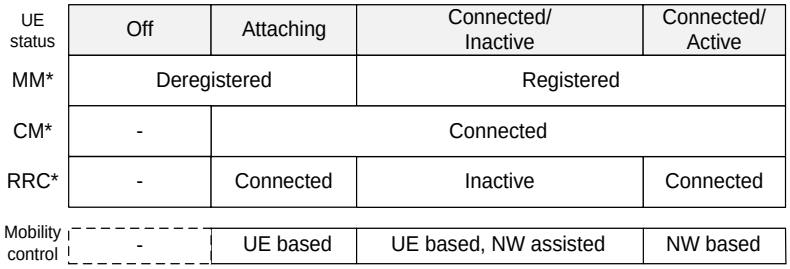


Figure 3-3: 5G Connectivity and UE RRC states

It is noted that the Figure 3-3 is mainly illustrating the RRC* always-connected approach and showing the relationship to core network Mobility Management (MM*) and Connection Management (CM*) states as an example. Figure 3-3 is not showing the RRC* Idle state, which is mainly intended for fault recovery and fallback procedures in case of link failure, see Figure 3-4.

3.4.1 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 the RAN perspective, the UE can be in RRC connected, RRC Connected Inactive or RRC idle state. When inactivity is detected, the UE may request based on a configured timer that the network can suspend the RRC connection. 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 3-4. The RRC suspend message may contain service tailored configuration in order to address UEs with diverse service requirements [SMS+16].

The UE state transition happens when either the UE or the 5G -NB (5G-NodeB) detects a low activity condition and UE has no ongoing data traffic in the user-plane.

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 sends 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 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.

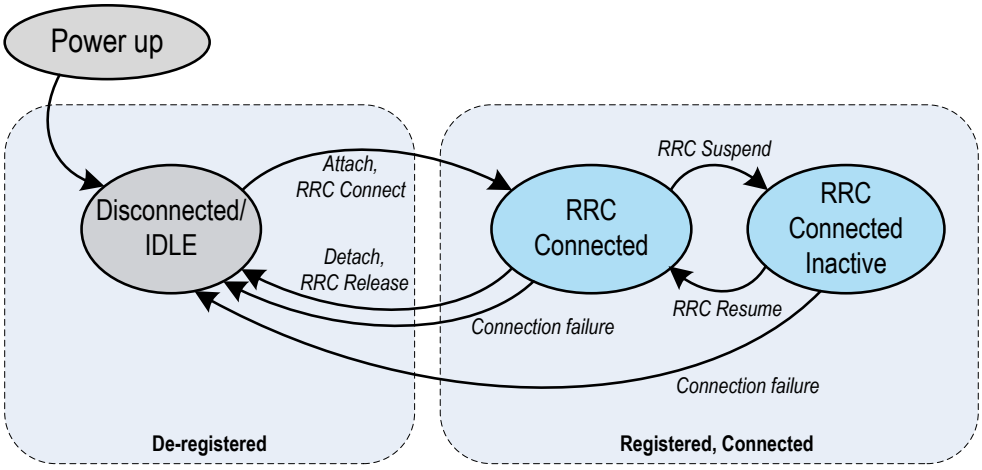


Figure 3-4: 5G UE RRC state transitions

3.4.2 State Transitions between Connected and Connected Inactive

Connection Suspend and Resume in Connected Inactive state

The Figure 3-5 illustrates the signaling flow of state transition from RRC Connected state to RRC Connected Inactive state and back to RRC Connected state. In this case, UE resumes connection to its Last Serving 5G-NB i.e. does not move while being in the Connected Inactive state.

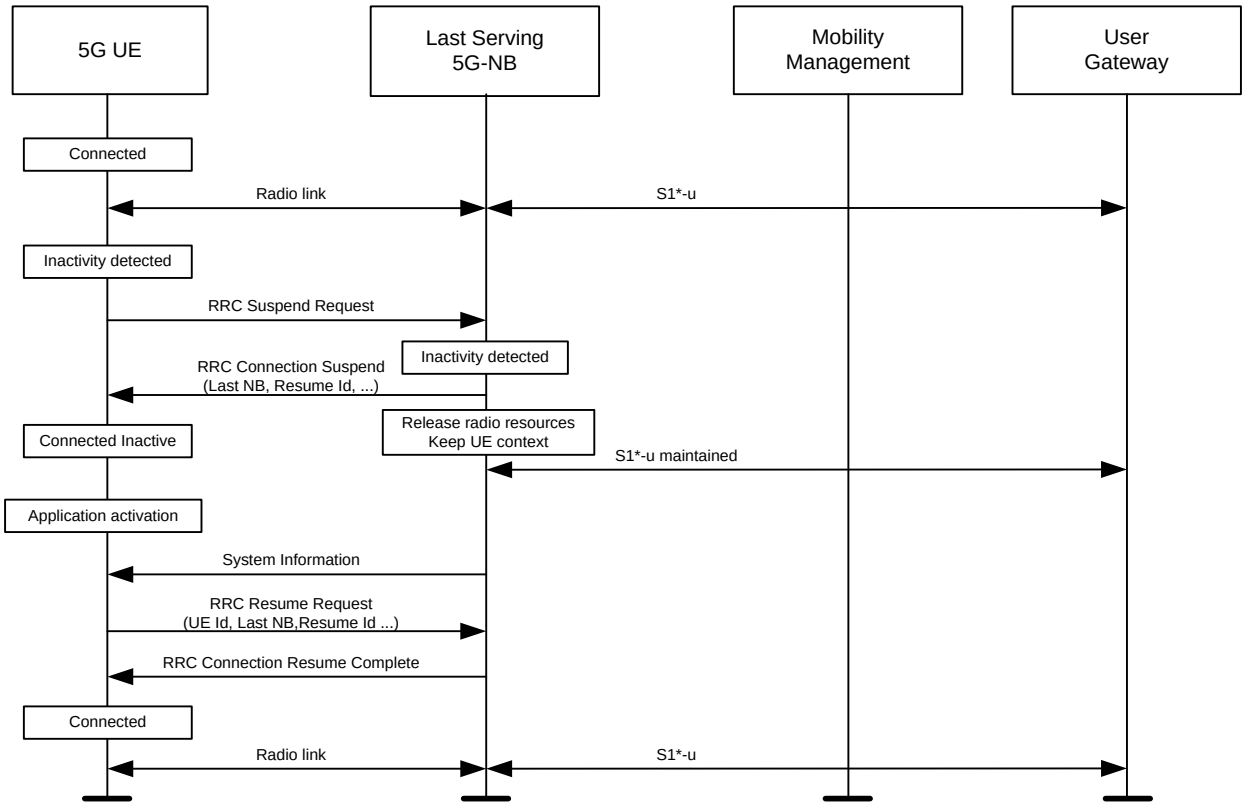


Figure 3-5, State transitions between Connected and Connected Inactive

The UE state transition happens when either the UE or the 5G-NB detects inactivity and UE has no pending data traffic in the user-plane buffers.

When 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 ID (in this case the Last 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.

Connectivity 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 selected cell and send 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 RRC CONNECTED state.

Network Originated Connection Resume in Connected Inactive state

Network originated connection establishment in Connected Inactive state in Figure 3-6 illustrates a scenario where UE has moved to other Cell/5G-NB within the allowed TA(s) while being in Connected Inactive state and the Last Serving 5G-NB has received registration over X2* signaling from UEs current location.

The Last Serving 5G-NB terminates the C-plane and U-plane connections (S1*) to the core network. When another peer (e.g. server) is sending DL packet(s) to the UE, these packets are received in Last Serving 5G-NB (RAN anchor) and are buffered there until UE has been reached by paging function from RAN.

The last serving 5G-NB is aware of UE's current selected 5G-NB based on registration and is able to transfer the required UE Context. Paging from all 5G-NBs within the UE's current allowed TA(s) is not needed. Upon reception of the UE context, the current 5G-NB takes the role of a new serving 5G-NB and resumes UE to its RRC Connected state.

Now the UE is in connected state and the forwarded DL packets can be delivered instantly when they arrive to new 5G-NB and after the patch switch UE has direct S1*-u connectivity to its serving User GW. The new serving 5G-NB sends a UE Location Update message to the Mobility Management (Chapter 5.3.2, Figure 5-10) to update the UE Context and location data.

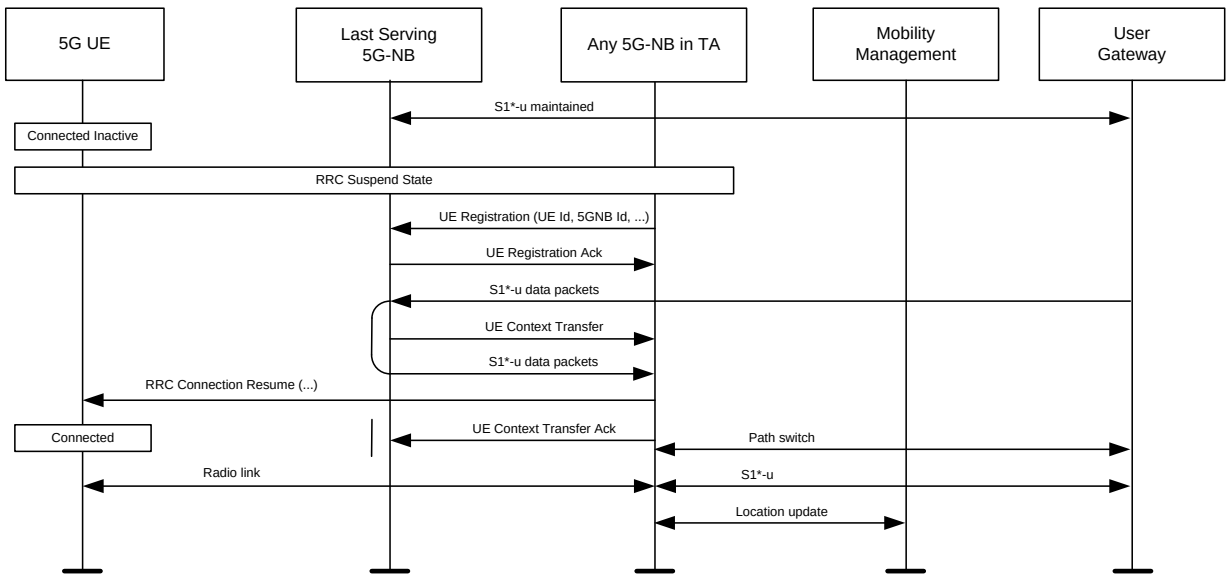


Figure 3-6, Network Originated connection in Connected Inactive

UE Originated Connection Resume in Connected Inactive state

The signalling flow in Figure 3-7 illustrates a scenario where an UE application wants to initiate data transmission and the UE has reselected to other Cell/5G-NB (other than Last Serving 5G-NB) within the allowed TA(s) while being in Connected Inactive state.

UE resumes the connection to the network via its current 5G-NB by sending a RRC Connection Resume Request message to the 5G-NB. In case a UE resumes the connection to an all new 5G-NB, the new 5G-NB should verify the UE Context from the Last Serving 5G-NB (RAN Anchor). New 5G-NB sends the UE NB Context Request message to Last Serving 5G-NB including at least UE ID.

The Last Serving 5G-NB verifies the UE and collects the full UE Context data to be included in the UE Context Transfer message that is then sent to the new 5G-NB with indication about the Serving NB change.

Upon reception of the UE NB Context Transfer message the new 5G-NB responds to the UE with the RRC Connection Resume Complete message to setup UE back to RRC CONNECTED state. UE is able to send its UL packets over radio link instantly when the U-plane resources over the radio link connection are configured and scheduled. Finally, the new serving 5G-NB sends a UE Location Update message to the Mobility Management to update the UE Context and location data.

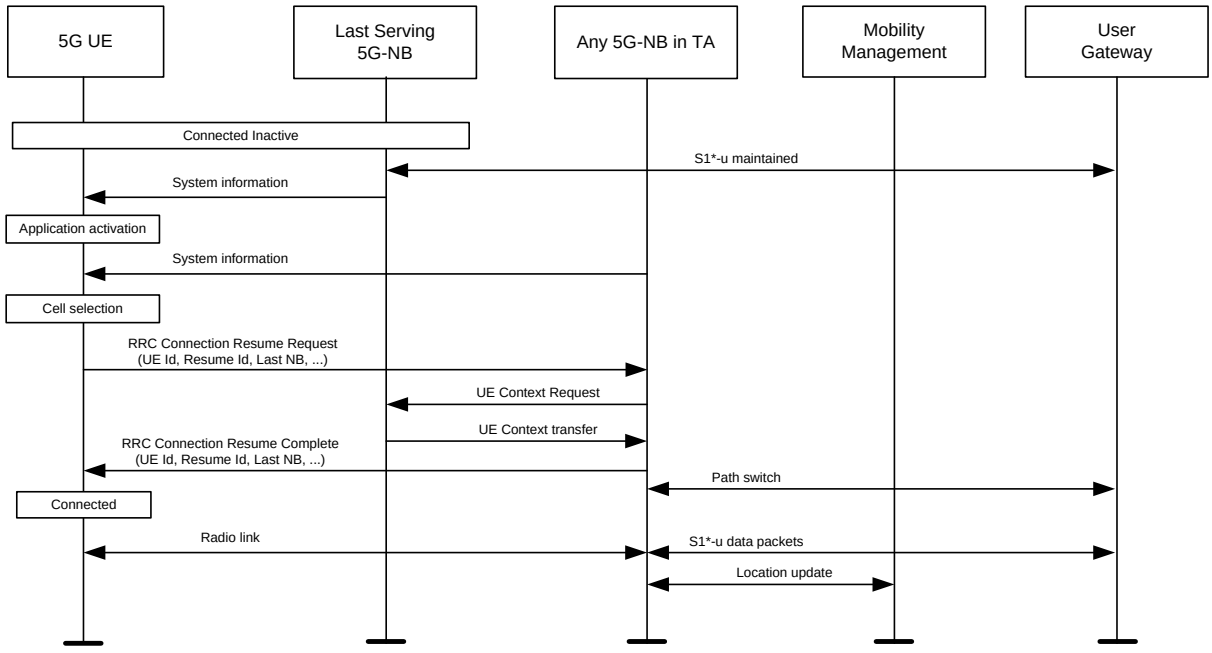


Figure 3-7, UE Originated connection in Connected Inactive

3.4.3 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 & 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 of vehicle, 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 of vehicle, 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 [3GPP15-22891] 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 [3GPP15-22891] 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 3-8, 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.

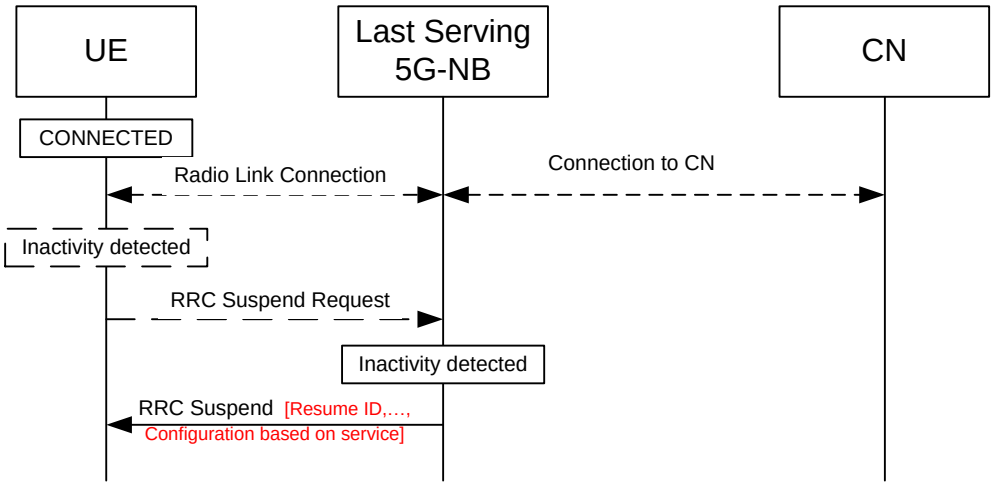


Figure 3-8: Configuration of a RRC Connected Inactive state

3.4.4 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 in order 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 comprises of packets arriving bursty to network and can be widely dispersed in time [3GPP11-36822]. In order to minimize UE state transitions and take into account the learnings from existing technologies, the MTC and IoT type of devices and applications would benefit from minimal signaling overhead to enable efficient short message transmission, and especially to minimize 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, lower 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 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.

3.4.5 The 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 the 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 signalling 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 mobility state design for 5G should learn from the pros and cons of state handling design in the existing technologies and take into account the 5G use cases and their requirements. Some of the benefits of the proposed state model that is:

- 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.

3.4.6 Impact on the Integration of LTE Evolution to the 5G

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 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 and UE is using the RRC Connected Inactive state during the low activity period. With tight integration the connection resumption back to RRC Connected could be done based on the system, which is able to provide better coverage or capacity. That is, a connection suspension from 5G may be followed by resumption in LTE when 5G coverage is not available in the same geographical area.

3.4.7 Impact on the 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, contention based uplink transmission and core network based location tracking.

4 Initial Access

4.1 Introduction

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 [3GPP15-36300]. This section presents mechanisms for certain areas that need particular enhancements and/or changes. Specifically:

- 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 layerRandom Access procedures addressing diverse access latency requirements and for a wide frequency range
- Paging optimizations for RRC Connected (Inactive) UEs.

Regarding the first aspect, new schemes aim to decouple system information distribution 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 beamform system information, proven very inefficient as shown in this chapter. A consequence of that is the usage of self-contained transmissions where system information decoding is not scrambled with the cell ID. Compared to the LTE, which is based on the SIBs and the MIBs, this proposal enables reduction of the energy consumption (since not all the nodes need to transmit) and flexible deployments allowing the physical split of UP / CP functions.

For the second aspect, the synchronization sequence(s) for coverage detection may be beamformed. In section 4.3.2 the benefits of digital BF in the initial access phase in terms of latency (consider fast transition from idle to connected state) and overhead are presented. Compared to analog BF, the delay is significantly reduced with the price of higher energy consumption. On the other hand, the increased energy consumption of digital BF can be offset by using low quantization ADCs, thus still enabling the use of the proposed scheme. Additionally, compared to the LTE's four-step initial access (i.e., comprising synchronization (Sync), random access (RA), connection request, and, contention resolution) the major difference of the designs is the fact that some of the steps can be done directionally. Alternatively, omnidirectional periodical transmissions may reduce the overhead of beamforming the broadcast information.

The third aspect handled in this chapter is related to the service differentiation in the initial access. In particular, the proposed solution focuses on handling the different delay requirements by using a combination of preambles. Compared to LTE, where only one preamble is used for the initial access, this proposal ensures the successful initial access for certain services. Additionally, the

proposed scheme reduces the retransmissions for all the services (both high and low priority ones) since it enables the simultaneous use of all the preambles by all the UEs.Finally, the fourth aspect is related to initial access based on Paging messages. The investigated aspect handles the Paging overheads when a UE is in RRC Connected Inactive state. Following the considered approach only the cell where the user camps is being paged, instead of the overall TAL which takes place in the LTE. This, reduces significantly the cost of the UE paging on the one hand, and since the UE context is in the RAN, also the delay is reduced due to the less interactions with the core network.

4.2 System information

4.2.1 Introduction

System information distribution is an RRC function in LTE; this is also assumed for the 5G RAN, although some cross-layer aspects may play an important role in the 5G RAN. In LTE, system information consists of very different types of information, access information, node specific information, system wide information, public warning system (PWS) information, etc.

METIS-II is investigating different aspects associated to the system information such as its content and the way it is distributed. Initial results have focused on the different mechanisms to deliver / acquire system information in such a way that:

- Energy consumption is kept to a minimum (e.g. not all the nodes need to transmit)
- Flexible deployments allowing the physical split of UP / CP functions are feasible.

4.2.2 System Information Distribution

System information in LTE is acquired through broadcasting a certain amount of information in each cell. In LTE is structured by means of System Information Blocks (SIBs), each of which contains a set of functionally-related parameters. The SIB types that have been defined include:

- The Master Information Block (MIB), which includes a limited number of the most frequently transmitted parameters which are essential for a UE's initial access to the network.
- System Information Block Type 1 (SIB1), which contains parameters needed to determine if a cell is suitable for cell selection, as well as information about the time-domain scheduling of the other SIBs.
- System Information Block Type 2 (SIB2), which includes common and shared channel information.
- SIB3–SIB8, which include parameters used to control intra-frequency, inter-frequency and inter-RAT cell reselection.

- SIB9, which is used to signal the name of a Home eNodeB (HeNB).
- SIB10–SIB12, which includes the Earthquake and Tsunami Warning Service (ETWS) notifications and Commercial Mobile Alert System (CMAS) warning messages.
- SIB13, which includes MBMS related control information.
- SIB14, which is used to configure Extended Access Class Barring.
- SIB15, which is used for convey MBMS mobility related information.
- SIB16, which is used to convey GPS-related information.

This list of System Information Block Types has been expanding over the years and is expected to keep increasing during the upcoming 3GPP releases.

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.

The time-domain scheduling of the SI messages (for the other SIBs) is flexible: each SI message is transmitted in a predefined periodic time window, while physical layer control signaling indicates in which subframes within this window the SI is actually scheduled. The scheduling windows of the different SI messages (referred to as SI-windows) are consecutive (i.e. there are neither overlaps nor gaps between them) and have a common length that is configurable. SI-windows can include subframes in which it is not possible to transmit SI messages, such as subframes used for SIB1, and subframes used for the uplink in TDD.

The physical channel which this information is transmitted on differs from block to block. For example, the MIB is transmitted over the Physical Broadcast Channel (PBCH) while the other SIBs are transmitted over the Physical Downlink Shared Channel (PDSCH) so they can be flexibly scheduled in other portions of the frequency band. That requires the usage of C-RSs for the system information decoding since it depends on PDCCH decoding.

When it comes to the amount of information, the MIB contains 14 information bits (with additional 10 spare bits for future use and a 16 bit CRC) while SIB1 and SIB12 may contain up to 1000 bits, which makes them expand over more than 6 RBs (although they may have a different coverage requirement, thus lower repetition rate).

These so called “always on signals” 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 SIB distribution relies on the aforementioned broadcasting of information and uses always on signals thus coming with a certain cost in terms of energy efficiency and spectrum resources. The overhead incurred by broadcasting access information for 5G AIVs is estimated in Table 4-2, based on the system information fields of LTE and the assumptions of Table 4-1:

Table 4-1 – assumptions for the overhead estimation

Carrier bandwidth	100 MHz
TTI	0.2 ms
AI periodicity	5 ms
Number of beams	224
Number of antenna ports per symbol	4
Number of subcarriers available for access information	492 in 1st symbol 600 in other symbols
Modulation	QPSK
Code rate	1/24
MIB periodicity	20 ms
SIB1 periodicity	80 ms
SIB2 periodicity	160 ms
MIB size	24 + 16 bits
SIB1 size	142 + 16 bits at minimum 350 + 16 bits at maximum
SIB 2 size	240 + 16 bits

Some remarks on the previous description and the respective analysis:

- Hybrid analog/digital beamforming is assumed, allowing a simultaneous transmission of 4 beams with the expense of reduced maximum downlink transmit power. Consequently 56 symbols (4 TTIs) are needed for a full sweep of one data symbol.
- The MIB and SIBs are assumed to be transmitted at the same periodicity as in LTE.
- The sizes of MIB, SIB1, and SIB2 are the minimum and maximum sizes of LTE SI based on TS 36.331. A typical size of MIB1 is smaller than the maximum due to shorter PLMN and scheduling info list.
- The first symbol of the AI block is assumed to contain beam specific reference signal(s) and synchronization signal(s) in addition to the system information.

Based on the assumptions of Table 4-1, the overheads for MIB, SIB1, and SIB2, are depicted in Table 4-2:

Table 4-2 – overhead estimation

SYNC+RS+MIB	4 TTIs	4 %

SIB1	16 TTIs	4 - 8 %
SIB2	24 TTIs	3 %

Consequently, the total overhead is in the range of 11-15 %. It should be noted that this result is subject to a number of uncertainties, such as:

- Deployment specific aspects such as the number of parallel beams and system bandwidth.
- System design aspects such as the transport format, AI periodicity, numerology, frame structure, and access information to be broadcasted.

Nevertheless, it can be seen that the overhead from access information broadcast has the potential of consuming a significant part of the system capacity. To reduce this overhead, some parameters (e.g. PLMN) might need to be signaled in a dedicated manner, possibly combined with more efficient encoding techniques.

4.2.3 System Information Distribution using Self Contained Transmissions

A possible way to fulfill the requirement for energy efficient information distribution is to provide the main part of the system information on a per-need basis, allowing reducing the amount of always broadcasted system information and only including information needed to access the system, with node specific and common system information delivered by dedicated transmission to the UE, see Figure 4-1.

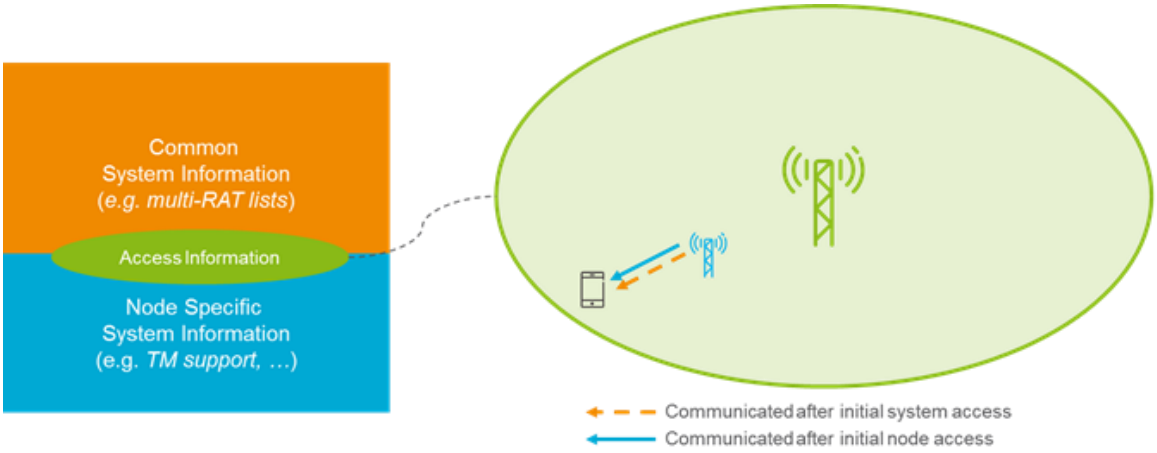


Figure 4-1 Access information distribution

The access information in the current assumption includes the random access parameters. These parameters include selected parts of the MIB, SIB1 and SIB2 information elements defined in LTE (e.g. PLMN Id, CSG, Q-RxLevelMin, Frequencybandindicator and Prach-configCommon). Energy efficiency may be achieved if one can define the previously described access information as a

self-contained system information signal i.e., with its own synchronization sequences, reference symbols and payload. By doing that the network does not need to transmit C-RSs over the whole band all the time (at least for the purpose of decoding system information). Future proofness can also be achieved by defining the access information as a self-contained signal once time-frequency resources are less occupied and allow the introduction of novel physical channels.

In the case of energy efficiency it is obvious that having signals broadcasted all the time and over the whole bandwidth disable the possibility of the network applying DTX cycles to its power amplifiers to save energy and control the level of interference being generated. This is potentially worse in ultra-dense scenarios expected in 5G. Therefore, in order to fulfill energy efficiency and future proofness requirements, METIS-II has started to investigate the impact of neighbors transmitting signals that follow an ultra-lean design of the broadcast system information, as shown in Figure 4-2.

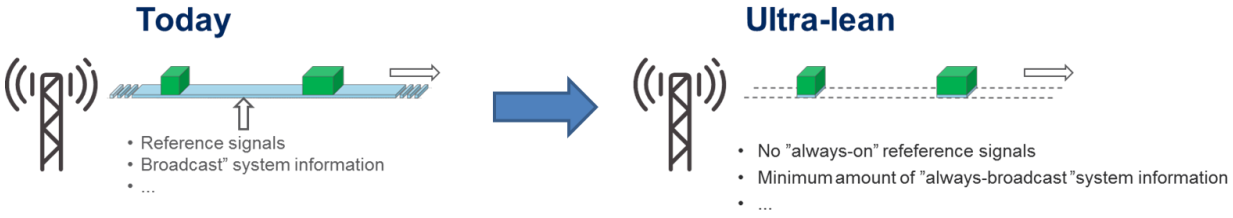


Figure 4-2 Ultra-lean design of reference signals and broadcast information

The adoption of the self-contained principle for access information creates further opportunities. The PSS/SSS would not be needed for decoding system information (i.e., since it is a signal with its own synchronization signals) so that the access information distribution does not need to be associated anymore with the cell concept (or any other network location concept). That principle would allow the split between system control plane and user plane (UP) where the UP access could have its own synchronization signals which would further enable the massive deployment of low-power access nodes without excessive overhead cost. In very dense deployments supporting very high data rates (e.g. by means of large bandwidth and/or a large number of antenna elements) the individual nodes will have no data to transmit or receive most of the time.

This separation from the cell concept could be beneficial in cases of unplanned small cell deployments, since the system information is decoupled from the cell concept where system information could be transmitted by overlaid nodes, jointly by MBSFN, or even in some cases (e.g., in cases of spotty 5G coverage) delivered by LTE Figure 4-7. That is possible since some of the nodes can have their own synchronization signals e.g. for dedicated transmissions.

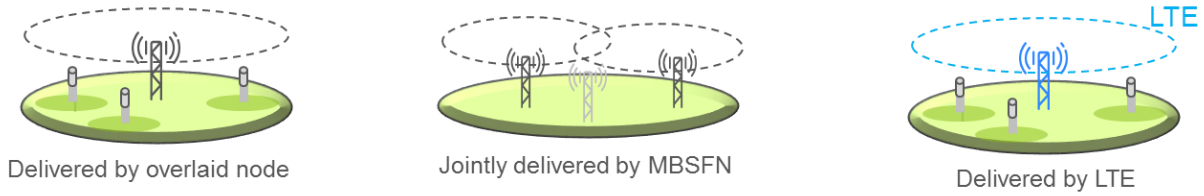


Figure 4-3 Flexible ways to deliver the access information

Another way to explore the fact that system information is the same among multiple nodes and at the same time reduce the amount of broadcasted information is by considering a different way to encode system information or, more specifically, the access information. The idea consists of an Access Information Table (AIT), containing a list of access information configurations and a short System Signature Index (SSI) which provides an index pointing to a certain configuration in the AIT, defining the access information, see Figure 4-4.

Figure 4-4 Transmission of AIT and SSI

In addition to that, the solution is also suitable for high frequency operation, which will massively rely on BF where broadcasting may become extremely inefficient in terms of wasted capacity. Therefore, the reduction of the existing system information to the minimized access information reduces always on signals.

The content of the AIT is assumed to be known by the UE when performing a random access attempt (e.g. during state transition). The AIT in the UE can be updated in two ways:

- A Common AIT (C-AIT) is broadcasted by the network, typically with a longer periodicity than the SSI e.g. every 500ms or so. In some deployments the C-AIT periodicity may be the same as the SSI periodicity (e.g. in small indoor networks) and the maximum C-AIT periodicity may be very large e.g. 10 seconds in order to support extremely power limited scenarios (e.g. off-grid solar powered base stations).
- A Dedicated AIT (D-AIT) transmitted to the UE using dedicated signaling in a dedicated beam after initial system access. The UE specific D-AIT may use the same SSIs to point to different configurations for different UEs. For instance, in the case of system congestion, this would allow to have different access persistency values for different UEs.

The SSI period is typically shorter than that of the C-AIT. The value is a tradeoff between system energy performance, UE energy performance and access latency in case SSI needs to be read before access. Different period configurations can be used to optimize energy consumption and to ensure UP and CP split.

Content of the AIT: One benefit with the idea is that the frequently transmitted SSI of limited size can be used to indicate the access information, signaled by C-AIT less frequently. C-AIT can also be transmitted on another carrier or received via LTE. This separation of the signals allows broadcasting the C-AIT on a longer time periodicity.

The AIT is envisioned to be available at the UE most of the time so that the UE would only need to acquire the SSI in most of the cases to acquire synchronization and, at the same time, assume the correct access information configuration before sending the random access preamble.

The efficiency of the described AI encoding method is dependent on the size (overhead) of the table relative to the geographical area covered by one table. In one extreme, all cells within certain area (e.g. metropolitan area) broadcast the same AI, hence allowing the AI to be encoded as a single entry covering that area. In the other extreme, all nodes within a certain area broadcast different AIs, the number of AIT entries being hence equal to the number of nodes within that area. Therefore, the efficiency of this method is subject to the reuse potential of 5G access information.

4.3 Coverage Detection and Synchronization

4.3.1 Properties of Synchronization Signals

The first step of accessing the LTE network is Cell Search [3GPP15-36300]. This procedure defines the time and frequency synchronization of a UE to a target LTE cell. To this end, the UE must detect the primary and secondary synchronization signals – abbreviated as PSS and SSS respectively. Specifically, PSS and SSS are transmitted together twice in a 10ms radio frame (every 5ms) occupying 6 resource blocks (i.e. 6x12=72 subcarriers). From a system access perspective, these signals encode together the Physical Cell ID (PCI) so the UE is said to be synchronizing with a cell. In addition to their logical meaning, another important aspect related to coverage detection is the way these signals are transmitted. In one alternative, these sequences are beamformed for the higher frequency regime (e.g., mmWave) In that case the UE should synchronize with a beam (an additional dimension: space, on top of time and frequency). In another alternative, these sequences are transmitted in an omnidirectional manner to minimize the overhead of beamforming. In the second case beamforming is only used in the step the UE receives a Random Access Request, where it is certain that there are users trying to access the network.

4.3.2 Synchronization Sequences and Random Access in Higher Frequencies

As described before, beamforming will further increase the challenges of initial access in general. Using multiple antenna elements to beamform, i.e., concentrate the signal energy in a specific direction, is the standard way to combat high pathloss in the mmWave frequencies and hence extend coverage. However, even with BF, due to low penetration of the signal in these high

frequencies the mmWave link will face acute and potentially frequent shadowing. Also, mobility in any axis (x,y,z) will cause the change of the spatial signatures of two ends of the link , leading to frequent SINR drops. These two factors make up one of the basic characteristics of the mmWave channel: intermittency.

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 the following issues are expected:

- We expect to have more quick drops of SINR due to channel intermittency;
- 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.

LTE's four-step initial access, comprises synchronization (Sync), random access (RA), connection request, and, contention resolution with the first two being the most challenging in a directional initial access procedure since:

- After Sync the UE must know the mmWBS position in the angular domain in order to establish the directional link to the mmWBS if beamforming is implemented on the UE side
- After RA, the mmWBS in turn must position the UE in the angular plane.

The major difference of the designs presented and evaluated to the current LTE initial access procedures, as shown in Figure 4-5, are:

- The two initial steps, Sync and RA, are done directionally,
- Sync and RA preambles are spread across four different frequency regions to gain frequency diversity. We assume a discretized 3D angular domain of maximum size NBS-by-NUE over which the mmW-BS and UE need to scan to detect the best angle pair. NBS and NUE are the number of antenna elements on the mmW-BS and the UE, respectively. If NBS = 64 and NUE = 16, the size of the angular space is 1024 angles.

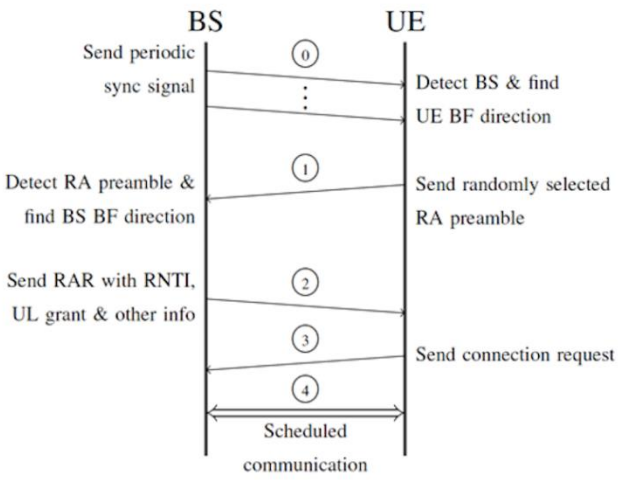


Figure 4-5 Directional initial access procedure

Figure 4-6 shows exemplary detection delays in milliseconds for a fixed overhead of 5%. The third and fourth column show the size of the discretized angular domain L, in sync and RACH for each option. K* is the number of full cycles required for detection in each scenario. Values for digital BF are derived given ADCs with 3 bit resolution. Synchronization/delay performance as function of overhead. Overhead is defined as signal duration over signal transmission period: T_{sig} / T_{per}.

Option	T_{sig}	L		1% SNR				5% SNR				High SNR			
		Sync	RACH	Sync		RACH		Sync		RACH		Sync		RACH	
				K^*	delay	K^*	delay	K^*	delay	K^*	delay	K^*	delay	K^*	delay
DDO	10 μs	1024	1	3	614.4	2393	478.6	1	204.8	65	13	1	204.8	1	0.2
	50 μs			1	1024	60	60	1	1024	3	3	1	1024	1	1
	100 μs			1	2048	17	34	1	2048	2	4	1	2048	1	2
DDD	10 μs	1024	64	3	614.4	24	307.2	1	204.8	2	25.6	1	204.8	1	12.8
	50 μs			1	1024	2	128	1	1024	1	64	1	1024	1	64
	100 μs			1	2048	1	128	1	2048	1	128	1	2048	1	128
ODD	10 μs	16	64	73	233.6	24	307.2	4	12.8	2	25.6	1	3.2	1	12.8
	50 μs			3	48	2	128	1	16	1	64	1	16	1	64
	100 μs			1	32	1	128	1	32	1	128	1	32	1	128
ODDig	10 μs	16	1	73	233.6	25	5	4	12.8	2	0.4	1	3.2	1	0.2
	50 μs			3	48	2	2	1	16	1	1	1	16	1	1
	100 μs			1	32	1	2	1	32	1	2	1	32	1	2
ODigDig	10 μs	1	1	79	15.8	25	5	4	0.8	2	0.4	1	0.2	1	0.2
	50 μs			4	4	2	2	1	1	1	1	1	1	1	1
	100 μs			2	4	1	2	1	2	1	2	1	2	1	2

Figure 4-6 Detection delays in milliseconds for a fixed overhead of 5%

Five basic designs based on the receiver's BF architecture (analog Vs digital) and transmission modes (directional Vs omnidirectional) are presented below. The following nomenclature is used: O stands for omnidirectional, D is analog directional, Dig instead means digital directional. The first letter is for mmW-BS Tx Sync mode, the second is for UE Rx Sync mode and, finally, the third is mmW-BS Rx RA preamble mode. Specifically:

a) DDO: Sync signal transmission and reception is done directionally in the analog domain. RA preamble is received omnidirectionally by the mmW-BS. The index of the time slot in which the UE received the sync signal is encoded in the RA preamble index. This way, the mmW-BS learns the TX direction in which the sync signal was received.

b) DDD: Sync is performed as in a) but RA preamble reception is done directionally in the analog domain. Since the mmW-BS scans the directions for the RA preamble, the BF direction can be learned from the direction in which the RA preamble is received.

c) ODD: Sync signal is transmitted omnidirectionally and received by the UE directionally. Here, RA is identical to b).

d) ODDig: Sync is as in c) but the RA preamble is received by the BS with a digital BF. Using digital BF enables the mmW-BS to have access to all the spatial samples at once, and hence learn the incoming direction in one measurement time.

e) ODigDig: Sync signal is transmitted omnidirectionally but is received at the UE with digital BF. Hence, the UE can learn the incoming direction at once. RA is identical to d).

Note that the size of the angular space changes based on the transmission mode.

For example, in Sync a) and b) the size is 1024; in Sync c) and d) it is 16, because only the UE scans the angular domain, while the mmW-BS sends the Sync signal in a fixed (omnidirectional) angle.

The detection of the Sync and RA signals is performed using matched filters. We need to stress that in the case of a cell edge UE, where the SNR is expected to be low, the angular space may need to be scanned more than once. This is proved to be very expensive in cases where the angular space is large, e.g., Sync in a) and b), as shown in Figure 4-6.

Figures Figure 4-7 and Figure 4-8 show the performance of the five designs, in Sync and RA respectively, as a function of the overhead at the cell edge. The superiority of digital BF is pretty clear.

To compensate for the increased power consumption that digital BF may entail, we propose the employment of analog to digital converters (ADCs) of low quantization ([BHR+15], [BHM+15], [BHM], [ZVM12], [RVM12]). This way, the energy consumption is brought at the same level (or even lower) as analog BF.

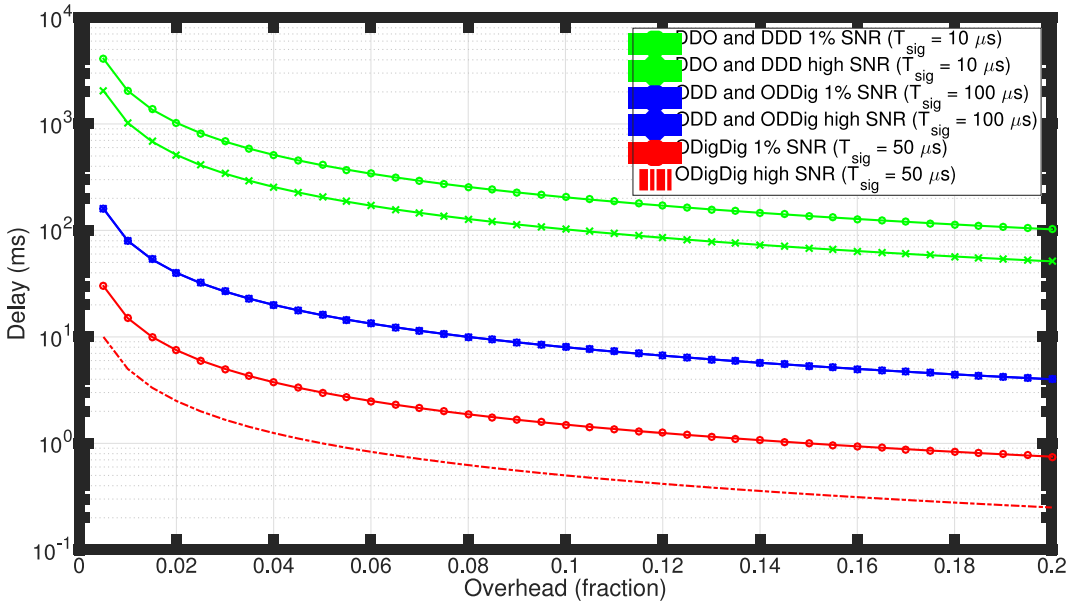


Figure 4-7: Synchronization delay performance of the designs as a function of overhead

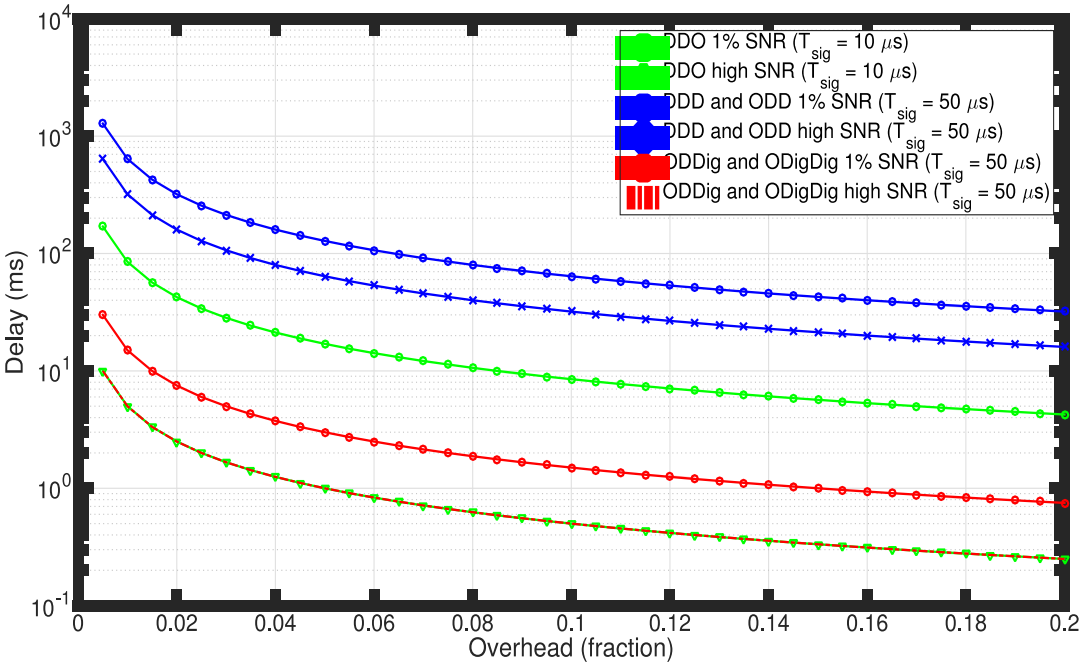


Figure 4-8: RACH delay performance of the designs as a function of overhead

The previous analysis leads to the outcome that, for the considered parameters, 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. 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.

Another alternative aims to reduce the overhead of beamforming the broadcast information. This is can be done by periodically transmit the broadcast information omnidirectionally. When a UE sends a preamble, the Random Access Response (RAR) can be beamformed.

Note that there is also the possibility to use a hybrid analog/digital beamforming approach, which is not investigated in this section. Hybrid beamforming may be considered as a compromise between the two extremes explored above, since it provides more signal detection opportunities compared to analog BF.

4.4 RACH Multiplexing in Support to Diverse Access Requirements

In LTE, during initial access, the UE randomly selects one of the random access preambles (64 preambles) informed via broadcasted system information. The current design may create problems e.g. for mMTC applications, where a large number of devices may simultaneously attempt to access the system (Figure 4-9). 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. These potential collisions will lead to additional access delays which may impact services differently.

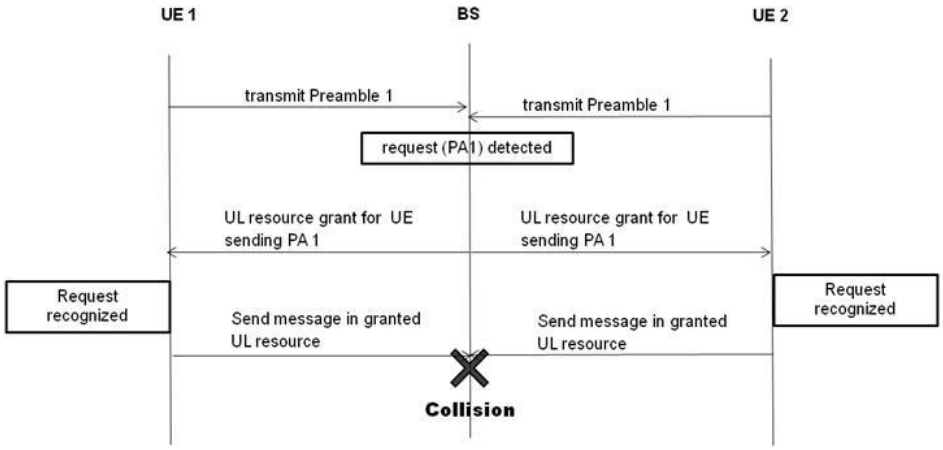


Figure 4-9: Collision in LTE random access.

Up to now, several schemes have been proposed in the literature for handling the random access procedure. These schemes may be classified into two large groups, namely pull-based and push-based [CKS+15]. In the first set of the solutions, the RACH preambles are being split in prioritization groups and the more delay sensitive devices compete against fewer devices for

accessing the system. Additionally, the time to have a second attempt to access the system could be fine-tuned according to the collision rate. In the second set of solutions, push based procedure is used to achieve Small Data Transmission (SDT) from the MTC devices. However, these schemes, are designed mainly for prioritizing access based on the transmission requirements and are not, on the one hand, aiming at solving 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 xMBB 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 large number of devices is considered (e.g., [MET13-D11], [NGM15]) highly questionable. Furthermore, the sparse RACH resources may increase the latency to unacceptable levels, for ultra-reliability scenarios.

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 more strict delay requirements to have higher priority, since combinations of preambles can always be identified by the receiver. This way, requests with higher priority are immune from collisions and the retransmissions (Figure 4-10).

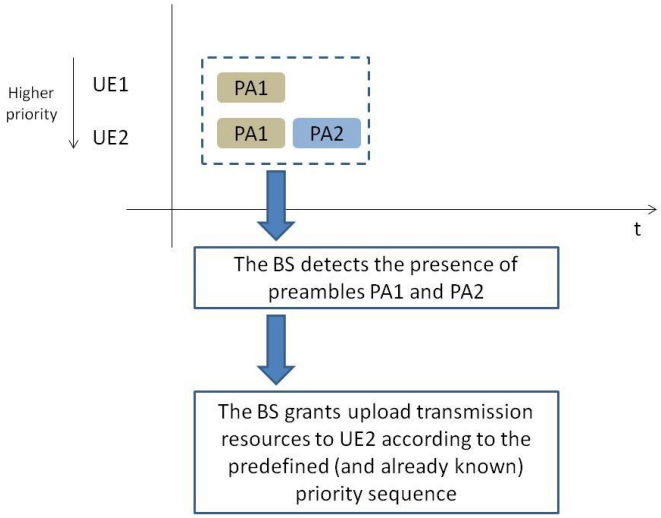


Figure 4-10: Preamble combination for prioritized UE

As it is shown in Figure 4-10, 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. Figure 4-11 illustrates the resolution procedure when two UEs with different priorities have collisions. 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.

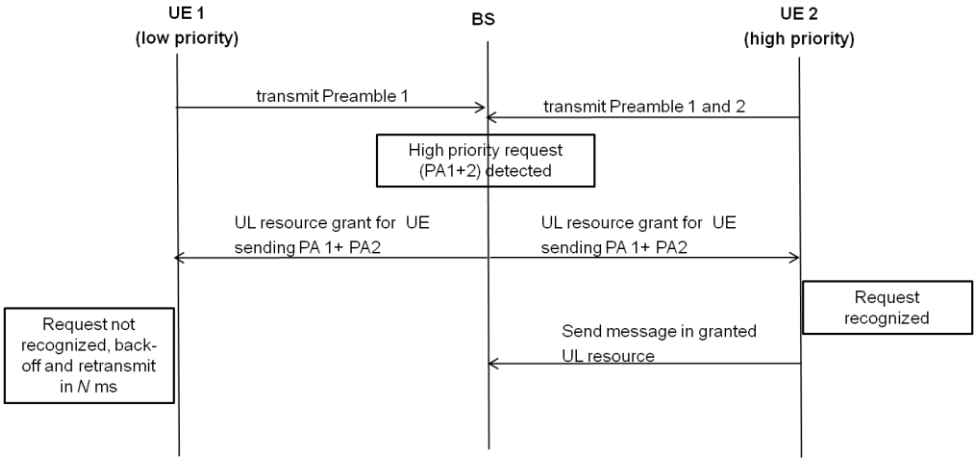


Figure 4-11: Collision in proposed random access scheme.

Similar approaches could be followed for more than 2 priority levels (Figure 4-12). In this case the device transmitting the higher number of preambles will have higher priority in the transmission request and will be granted the resource grant, whereas the other two UEs (i.e., UE1 and UE2 in Figure 4-12) will not take the transmission grant.

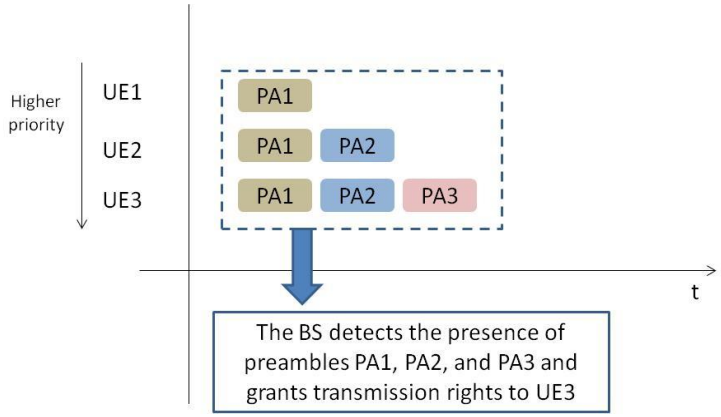


Figure 4-12: Preamble combination for 3 priority levels.

The combined preambles per UE may be multiplexed in time, frequency, or even both domains. The combination of preambles, as well as the time/frequency shifts Δt and Δf may be predefined or determined dynamically (by considering the current load conditions) and communicated to the UEs via system information. The RAN utilizes this knowledge on the determined time and frequency shifts to avoid the possible misdetection.

The overall description is schematically provided in Figure 4-13. As depicted, the UEs receive via system information the preamble combinations linked to the corresponding priorities. The preambles combinations function may dynamically re-calculate the preamble priorities depending on the load conditions (and in general the network status) or network policies per service.

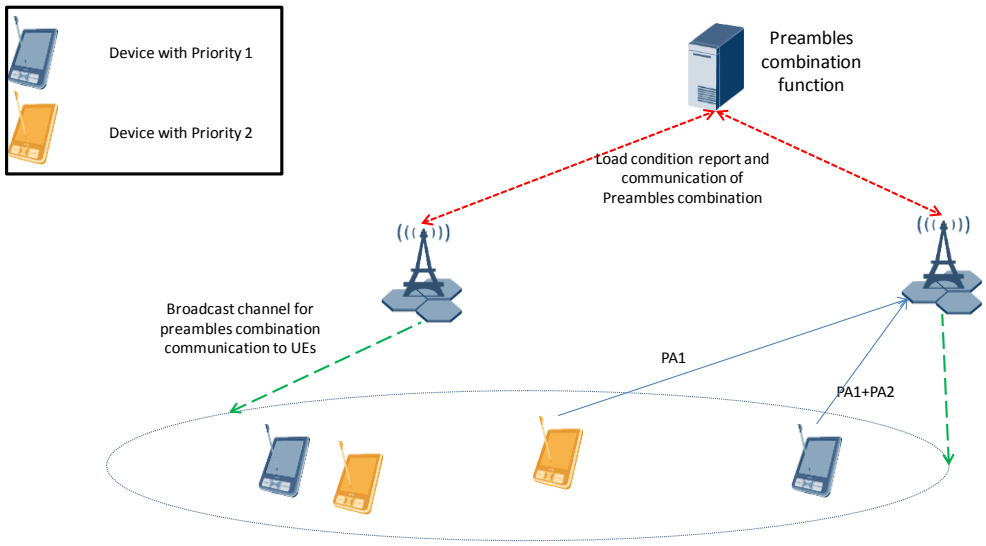


Figure 4-13: Schematic representation of the preambles combination distribution and use scheme.

Other methods for random access prioritization per service could also be considered in the future within METIS-II such as the adaptation of power settings per service.

The proposed scheme has been evaluated in terms of the retransmission need, compared to an access class barring scheme based on LTE. Specifically, the state of the art solutions belonging in this category use transmission of single LTE preambles for system access; the preambles are being separated into groups of different priority categories. The timeslots and the preambles are statically allocated to each class/group [CKS+15].

In this comparison we consider two priority levels. Out of the total number of system access requests, the 10% is considered to have higher priority. Additionally, we assume that we have 50 available preamble sequences, which could be used for initial system access.

The state of the art solution is an access class barring scheme where the preambles are being statically split into two groups. In the higher priority system access requests we allocate 20 preambles, whereas in the lower priority system access requests we allocate the 30 preambles. It is assumed that the allocation of 20 preambles out of the total 50 available ones could be

reasonable for covering the high priority requests which is the 10% out of the overall number of requests. For the proposed solution, as described above, we assume two types of system access requests for the two different priorities. The devices that transmit lower priority requests use only one preamble signature, whereas the devices that transmit higher priority requests send two preambles simultaneously; the number of the considered preambles is the same as in the access class barring scheme.

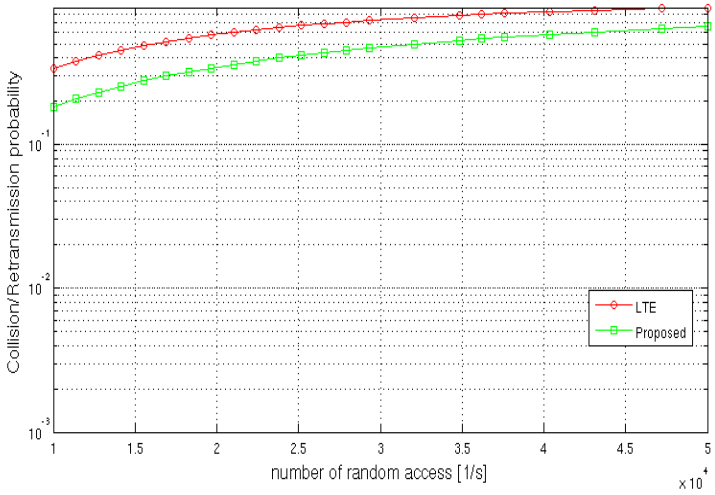


Figure 4-14: Comparison of collision/retransmission probability for low priority (level 2) request

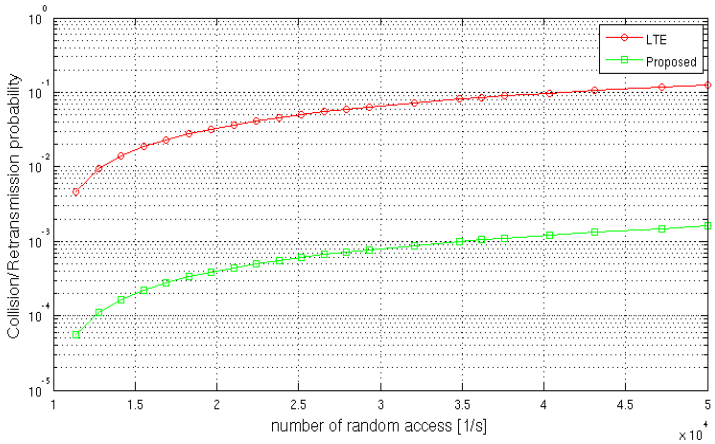


Figure 4-15: Comparison of collision/retransmission probability for high priority (level 1) request

Figure 4-14 and Figure 4-15 provide the collision/retransmission probability for the low and high priority requests. In both schemes (the proposed one and the access class barring scheme) we observe that the high priority requests have significantly lower collision probability than in the low priority requests. In the state of the art solution, this is quite reasonable since we reserve a big

amount of preambles (40%) for a certain portion (10%) of the overall requests. In the proposed scheme the high priority devices use the combination of preambles which in case of collision it overwrites the preamble of the lower probability request. We observe that collisions still occur even for the high priority requests, but with very low probability (~0.2% when we have 50000 requests/sec). In any case, we observe that the collision probability is much lower compared to the state of the art solution (~100 times) even though a huge amount of preambles is reserved.

Another interesting outcome of this analysis is that for the low priority requests the proposed scheme outperforms the access class barring scheme. The reason is that in the latter, the devices with the low priority requests compete for a significantly smaller amount of resources (i.e., 30 instead of 50 preambles), which leads to higher collision rate. Specifically, even when in the state of the art solution there is practically 100% collision rate probability (when the amount of requests per second is very high) in our proposal the device may still access the system.

4.5 Paging

4.5.1 Introduction

Paging is a system access functionality that enables the network to initiate a contact to registered UEs. In LTE, paging is initiated by the CN to contact an RRC Idle UE, e.g. when DL data is detected for the UE. The location of an RRC Idle UE is known by the CN, i.e. MME, on a TAL level where a TA in LTE is a group of cells. The TAL is given to the UE when it registers to the network, e.g., through *Attach* procedure and updated periodically or through a trigger, e.g. when the UE moves to a new TA that is not in its TAL. The mobility of a UE within its registered area is autonomously controlled by the UE.

In order for the network to be able to reach the UE in RRC Idle state, the UE has to monitor the paging channel of the cell it has camped on. The monitoring of the paging channel is typically done in a discontinuous manner using a network configured DRX cycle in order to prolong UE's battery life. When the network wants to reach the UE, it pages the UE through all the cells in the TA in the TAL to which the UE is registered. When a cell receives a paging message via S1 signaling to be broadcasted to a UE, it waits until the next Paging Occasion to send the paging message to the UE. Most of the paging messages broadcasted by network are, however, a vain attempt to locate UEs as TAL typically consists of a large number of cells.

In LTE, the paging load on the AI and the CN / RAN interface is significant when compared to other signaling loads. In average loaded network, MME can experience a sustained signaling load from 500 to 800 messages per UE during a normal peak busy hour and up to 1500 messages per user per hour under adverse conditions [ALC13]. According to [ALC13], paging accounts for 29% percent of the total signaling load at the MME, while TA updates (TAU) represents 5% of the total load. Since a single S1-paging message that is sent from the CN to RAN is typically broadcasted across a large number of cells, the paging load on the AI is much higher than the load on the CN/RAN interface.

With the expected massive number of devices and more small cells (e.g. due to higher frequencies and denser deployments) in the 5G time frame, paging may significantly increase the load on both the AI and the CN/RAN interface. Tackling the problem by simply shrinking the size of TAL may not work as the signaling load due to TA updates will increase significantly. Initial analyses in WP 6 conclude that efficient paging and UE location tracking concepts need to be addressed in the 5G RAN design. Several paging schemes have been discussed in the academia and industry with the objective of reducing the paging load. In *sequential paging*, cells are paged sequentially which may lead to a substantial increase in the delay to reach a UE. In *adaptive paging*, cells are paged sequentially but in a smart manner, such as paging the last known position of the UE first, or paging randomly in a parallel manner, etc. [Chu11], [XCG06]. However, the adaptive paging schemes may lead to an increase in the delay for reaching UEs. On the other hand, an adaptive paging scheme based on sending *"information on recommended cells and eNBs for paging"* by the last serving eNB to the MME has been discussed for paging optimization for LTE Release 13 with the objective of reducing the paging load on the AI and the processing load on the MME [3GPP15-150698]. However, such scheme may lead to an increase of the delay for reaching UEs as there is a probability that the UE may not be reached during the first or more paging attempts [3GPP15-150946].

4.5.2 Paging Design and Directions

METIS-II is currently investigating one approach to achieve efficient UE tracking using a hierarchical location tracking where the CN tracks the registration of a UE on a group of RAN locations and the RAN tracks a higher granularity of the UE location, e.g. something equivalent to a cell-level location for UEs in RRC Connected (Inactive) state.

The *RAN location* of the UE could possibly be updated to the RAN for RRC Connected (Inactive) UEs. This update could be done by using a light signaling procedure which is terminated in the RAN. The details of such a light signaling procedure will be investigated along the security handling mechanism for the messages exchanged in the signaling procedure, for example, by retaining and updating the security context from the last attach procedure. Using the considered hierarchical location tracking approach, the paging and part of the mobility functionalities of the CN would be moved to the RAN. In such cases, the UE will be paged only in the cell where it resides or based on its later update of its last position to the RAN based mobility function (Figure 4-16) compared to the LTE where all the Tracking Area would be paged.

The RAN-based paging and location tracking has superior performance in terms of signaling overhead and mobile terminated packet latency when compared to the CN-based approach. Based on the message count, e.g. from Figure 4-16, there is at least 84% overall paging load reduction. Further reduction can be obtained using optimizations such as clustering of UEs for paging and location update. On the other hand, the latency for 1st mobile terminated packet transmission in the RAN-based approach is much lower than in the traditional LTE case. Specifically, it is lower than LTE by the amount of:

$$2^{*} \text{ radio RTT} + 1.5^{*}\text{S1C RTT} + \text{S11 RTT}$$

Where, RTT represents the Round-Trip Time of the protocol.

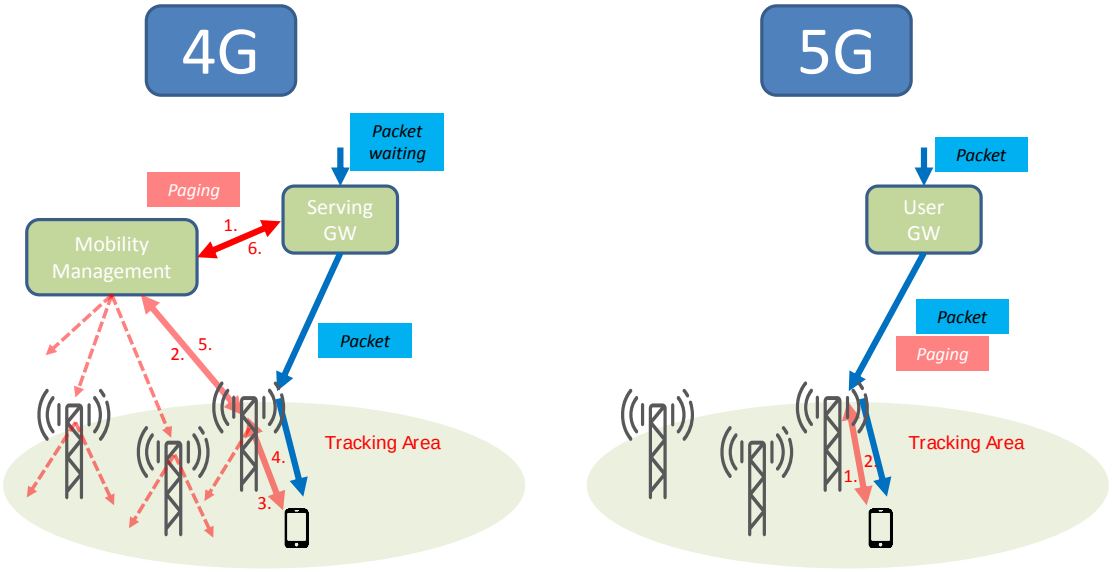


Figure 4-16: Paging solutions in a) LTE pre Release 14 and b) 5G

One of the advantages of such an approach is the potential to significantly reduce the paging load on the AI using a higher granularity paging, especially in the case of semi-static UEs where these *RAN updates* would not be so frequent (and basically the network gets the UE location for free). The network can use the RAN level location of UEs to proactively forward data to the 5G-NB where a UE is known to be camping.

For high mobility UEs (especially assuming small cell deployments) the approach may create some drawbacks in terms of the location update signaling generated by the moving UEs when performing cell reselection in RRC Connected Inactive state and increased power consumption. In this case, METIS-II envisions further investigating the impact of different mobility patterns for that approach.

Another advantage of the proposed scheme is the potential of significantly reducing the paging load on the CN/RAN interface. Assuming that in Connected (Inactive) state the 5G RAN keeps the S1* connection alive, the CN is not aware whether the UE is active or inactive. For UEs in that state, the CN just forwards incoming packets to a given S1* termination for that UE at a given RAN logical node (denoted herein 5G-eNB). As stated earlier, the scheme is optimized for the case the UE is semi-static and remains in a certain location controlled by the same 5G-eNB. When the UE moves and notifies a new 5G-eNB there has to be some mechanism that allows the CN to forward packets to the right RAN termination point. If nothing is done the CN will just forward to the latest RAN node the UE was active in. That last active 5G-eNB could also act as mobility anchor point but would need to be aware of the new UE RAN location (and forward packets via the X2* interface) or leave the mobility anchor responsibility to the new RAN node the UE has moved to.

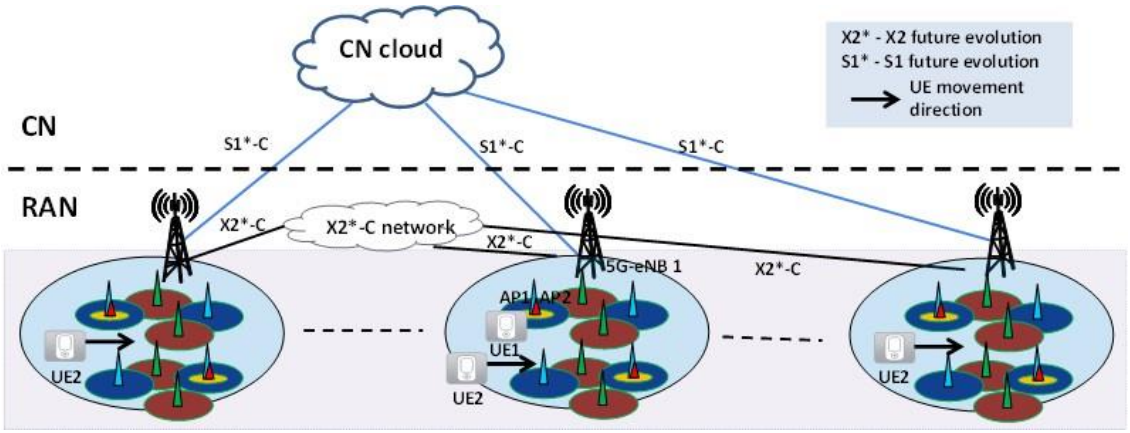


Figure 4-17: A heterogeneous network with dense small cell networks which may operate on different carrier frequencies. Two RAN areas working on different carrier frequency may become co-located.

As an example of how the hierarchical location tracking can be implemented, consider a heterogeneous network setting shown in Figure 4-17. In this example, for simplicity, let us assume a minimum RAN-level location is a *cell*, as in LTE. Assume 5G-eNB 1 terminates the S1* connection for the last serving cell of UE 1 and is serving as the RAN mobility anchor for UE 1 after it enters RRC Connected (Inactive) state. 5G-eNB 1 may track the cell-level location of UE 1, for example, by associating UE 1 to AP 1. Notice that the cell level location information tracked by the RAN can be generalized to a group of cell level location information, e.g. to minimize signaling load due to RAN level location updates when the UE is moving or for the RAN to be aware of the cells that provide the UE with the strongest average received signal quality on different carrier frequencies. For example, for the latter case, 5G-eNB 1 may associate UE 1 to AP 1 and AP 2 as two cells with strongest average received signal quality on different carrier frequencies. When DL data is available for UE 1, the CN forwards the data to 5G-eNB 1 and 5G-eNB 1 pages UE 1 only in the cell where it is known to be camping. This way, RAN is able to hide the mobility of UEs in connected inactive state between cells that are located in the same control area of a RAN mobility anchor, while also achieving high granularity of tracking the UE locations. The control area of a RAN mobility anchor may contain only the cells whose S1* connection is terminated in the RAN mobility anchor or it may also include cells whose S1* connection is not terminated in the RAN mobility anchor but could be reached using X2* interface.

When a UE reselects a cell which is not in the control area of the RAN mobility anchor, two options are considered for managing the UE's RAN context and S1* connection. One option is to keep the RAN mobility anchor functionality of the UE in the entity which terminated the S1* connection for the last serving cell of the UE and the switching of the S1* connection is done only when it is needed and a new DL data is available for the UE. Another option is to move the mobility anchor functionality to a new RAN entity that terminates the S1* connection for the new cell. In this case, the RAN informs the CN the identity of the new RAN mobility anchor for the UE, the S1* connection needs to be switched to the new mobility anchor and UE's RAN context is transferred to the new RAN mobility anchor. For example, assume in Figure 4-17 the RAN mobility

functionality is located in 5G-eNBs and the control area of the RAN mobility anchor includes the cells whose S1* connection is terminated in the 5G-eNBs. When UE 2 moves from one cell in the control area of one 5G-eNB to another 5G-eNB, the RAN mobility anchor functionality and S1* connection are moved to the new 5G-eNB. On the other hand, if the UE moves to a cell that is not in its registration area, it has to at first update its registration to the CN before continuing with the RAN level location update.

In LTE the MME initiates the paging procedure by sending PAGING message to each eNB with cells belonging to the TA(s) where the UE is registered [3GPP15-36300]. All the eNBs who receive the PAGING message perform paging of the UE in cells, which belong to tracking areas. The eNB who receives the Paging Response forwards it to MME, which informs the UE location to S-GW. Now S-GW can forward the UP data to correct eNB.

The benefit of RAN based paging over LTE is the possibility to minimize the latency between paging and initial data transmission. Using the proposed RRC state model for 5G, the UEs are in RRC Connected Inactive state, the UE context is in RAN and the RAN knows UE location for fast data forwarding of the first packet to the target cell. Further, when RAN based paging is used, there is no need to buffer the initial data at the User GW until the location of the UE is known with the granularity of a single cell or a cluster of small cells or access points under the 5G-NB. This approach can significantly reduce amount of paging messages over the AIVs and remove or reduce the data forwarding latency from User GW to base station for non-ideal base station backhaul connection such as S1* or X2*.

RAN based paging is suitable for low latency small data transmission, where the signaling time compared to the transmission time of small data is relatively long. Low latency for initial data can be considered a significant benefit since the volume of MTC or IoT devices is expected to be very high and core network procedures need to be optimized for resource efficiency. Considering the stationary MTC devices, which may be operating out-/in-doors, such a MTC UE may be connected potentially to one or more a small cells with the master cell being the macro cell. Therefore it is possible to proactively forward the data at least to the macro cell, which can do the paging and has the initial data available for transmission after the paging response from UE.

5 Mobility

5.1 Introduction

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 in RRC Connected (see Section 5.2) and low active users in RRC Connected Inactive state (see Section 5.3). In addition, it presents new ideas on how to explore the usage of context awareness and data analytics to optimized mobility algorithms, see Section 5.4.

5.2 Connected Mode Mobility

5.2.1 Introduction

As said earlier, the reliability and interruption delay requirements for of 5G are more stringent compared to 4G. In addition to this, 5G is expected to operate in a wider range of frequencies (1-100 GHz) than 4G. This also means that beamforming techniques may be needed to compensate for the higher propagation loss at high frequencies.

The support of beam forming mobility, efficient UL and DL measurements and the lean design of the reference symbols are important aspects for 5G. This also means that the 5G cell concept will probably not be same as in LTE. However, METIS-II still believes there will be a mobility function similar to LTE regardless of the cell concept and DL reference signal design in 5G, see Section 5.2.3. Exactly how this should look like will be further investigated. Due to the lean design of the DL reference symbols, the UL measurements can complement the DL measurements for beam forming mobility. One method how to perform UL measurements for beam forming mobility is described in Section 5.2.4.

How to perform the handover procedure to handle to very strict requirements is investigated in Section 5.2.6. It concluded that the handover procedure should follow the “make-before-break” handover procedures.

The beam forming mobility design is investigated in Section 5.2.7. The beam forming mobility design should support a fast switching/tracking of the communication beam to combat rapid changes in link quality. Also, the design should be able to exploit the availability of multiple overlapping beams that can be used for the communication with a single UE. Further on, the beam mobility should have a minimum impact to the RRC layer. One solution to fulfill these requirements is the idea of cluster based mobility, which is a set of nodes[5] that the UE can detect and which are prepared in advance for a fast re-routing of the signaling and user data. Fast re-routing is achieved at the expense of increased inter-node signaling and monitoring of downlink control channels.

[5] Note, as said earlier, that the cell concept is not clearly defined here and may very well be different from the LTE cell concept, and sometimes we instead use the more general term node or eNB instead of cell.

5.2.2 Connected Mode Mobility Related Functions

One of the main goals of the RAN mobility for the RRC Connected (Active) state where data is transmitted is to provide seamless service continuity across the network. In LTE, this is achieved via a set of procedures between the UE and the network and between source and target network nodes (e.g. via X2/S1 interfaces). METIS-II envisions that the mobility design for the 5G RAN in Connected (Active) state needs to address the following mobility-related functions:

- Measurements
  - DL based measurements
  - UL based measurements
- Measurement reporting
- Mobility triggering/decision
- Beamforming mobility

Measurement reporting and mobility triggering are 5G functions that probably will be similar to LTE. The new functions that will impact the mobility most are the DL measurements due to the lean design of the control signaling and the beamforming mobility.

It is initially assumed that the Connected (Active) mode mobility will, as in LTE, be controlled by the 5G RAN according to the UE activity, measurements and mobility. That is, the RAN decides when to make the handover and what will be the target. The handovers are based on the UE measurements and measurement reporting is controlled by parameters given by RAN. The CN/RAN S1* connection (see Section 2.3.2) is updated after the radio handover has been completed.

5.2.3 DL Signals to Support Mobility

In LTE, the UEs monitor the quality of the active connection using the C-RSs as well as the *UE-specific RSs*. These signals can be used for estimating e.g. the Channel Quality Indicator (CQI), the RSRP (Reference Signal Received Power) and RSRQ (Reference Signal Received Quality), see [3GPP14-36214] for more details. To measure the quality of neighboring cells, the C-RS are used. The *UE-specific RSs* signals may be embedded in the data for specific UEs and are often called DeModulation Reference Signals (DM-RSs). They were standardized in Release 8, but enhanced in Releases 9 and 10. Discovery reference signals (DRS) were introduced in Release-12, allowing more efficient small cell on/off operation via configurable periodicity of DRS occasions.

As said above, METIS-II believes the major difference between 5G and LTE for mobility will be the lean design of the control signaling and the use of massive beam forming in some scenarios. As said earlier, the reason for the lean design of the control signaling such as the reference signals is to be more energy efficient than LTE and also future proofness requirements, see Section 4.2. This means that 5G probably will not have the same cell concept as LTE with a C-RS transmitted continuously (see Section 4.2). However, METIS-II still believes there will a mobility function

similar to LTE regardless of the cell concept and DL reference signal design in 5G. Exactly how this should look like will be further investigated.

The other aspect that affects the mobility is that massive beamforming may be crucial for 5G in some scenarios. In order to handle this issue, different aspects should be considered. A first aspect that needs to be studied is the design of reference signals for the beams and its configurations used for link quality monitoring (for the serving beam) and neighbor link measurements and UE reporting mechanisms. Differently from LTE, these reference signals need to be designed to possibly be beamformed in wide and/or narrow beams. Assuming a network based approach as a baseline (as in LTE) the UE should be capable to synchronize and detect this signal(s) from neighbor beams(s) and report some indication to the network (e.g. absolute or relative neighbor beam quality/measurements) in order to get some sort of mobility command from the serving beam to the UE so the UE can synchronize and access the target selected beam before the sudden SINR drop occurs. The delay of this kind of procedure should be studied and alternatives should be considered (e.g. UL-based measurement procedures relying on UL sounding signals) for both intra-node beam mobility and inter-node beam mobility.

5.2.4 UL Signals for Mobility Measurements

Unlike in traditional LTE networks, the system proposed in [GMR+16] is based on the channel quality of uplink (UL) rather than downlink (DL) signals. This has several key benefits. First, the use of UL signals eliminates the need for the UE to send measurement reports back to the network and thereby removes a point of failure in the control signaling path. Second, if digital beamforming or beamforming with multiple analog streams is available at the mmWave node, then the directional scan time can be dramatically reduced when using UL-based measurements. Finally, since the base station is less power constrained than a mobile device, digital or hybrid beamforming will likely be more feasible at the BS side.

Therefore, the work in [GMR+16] proposes a novel multi-cell measurement reporting system where each UE directionally broadcasts a sounding reference signal (SRS) in a time-varying direction that continuously sweeps the angular space. Each potential serving node scans all its angular directions and monitors the strength of the received SRS along with its variance, to better capture the dynamics of the channel. A centralized controller obtains complete directional knowledge from all the potential nodes in the network to make the optimal serving node selection and scheduling decision.

In the proposed framework, illustrated in Figure 5-1, there is one major node called MCell (Master Cell, in accordance with 3GPP LTE terminology), which is typically a microwave base station. However, functionally, the MCell can be any network entity that performs centralized handover and scheduling decisions. The UE may receive data from a number of nodes, either mmWave or microwave, and we call each such node an SCell (Secondary Cell). In order to communicate and exchange control information, the SCells and the MCell are inter-connected via traditional backhaul X2 interface connections, while each user can be reached by its serving MCell through the legacy 4G-LTE band.

The network will monitor the signal strength on each of the directions pairs for each of the possible links. This is done by each SCell building a report table (RT), based on the channel quality of each receiving direction, per user.

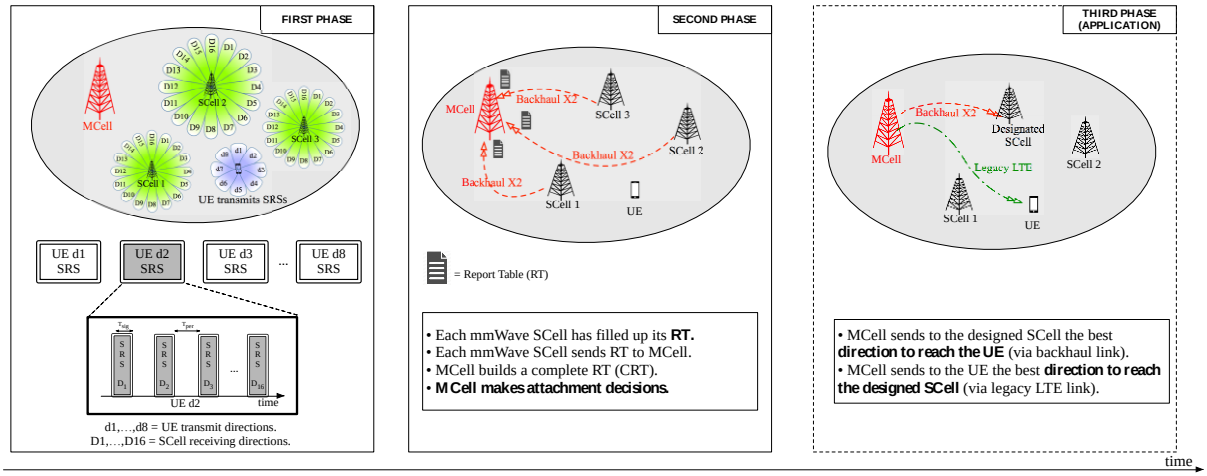


Figure 5-1: Slot scheme for the proposed UL-based procedure

1) Phase 1: Uplink measurements: The UE directionally broadcasts uplink sounding reference signals in dedicated slots, steering through directions 1, . . . , N_{UE} , one at a time, to cover the whole angular space. The SRSs are scrambled by locally unique identifiers (e.g., C-RNTI) that are known to the SCells. Each SCell performs an exhaustive search, scanning through N_{BS} directions, in order to fill the i^{th} row of the report table, which refers to the user steering direction i. The quantity:

$$SINR_{i,j} = \max SINR_{i,j}(k) \ (1) \ _{k=1,\ldots,N_{BS}}$$

represents the highest perceived SINR between the UE, transmitting through direction i, and $SCell_j$, maximized over all its possible receiving directions. The value:

$$d_{i,j} = d(SINR_{i,j}) = d \max SINR_{i,j}(k) \ (2) \ _{k=1,\ldots,N_{BS}}$$

is the angular direction through which such $SINR_{i,j}$ was received by $SCell_j$.

2) Phase 2: Network decision: Once the RT of each SCell has been filled, each mmWave cell sends this information to the supervising microwave MCell through the backhaul link which, in turn, builds a complete report table (CRT), as depicted in Table 5-1. When accessing the CRT, the MCell selects the best mmWave BS candidate for the considered user, based on different metrics. For example, the MCell could select the maximum SINR (with some hysteresis), in order to have the best channel propagation conditions. Such maximum SINR is associated, in the CRT's entry, to the SCell direction d_{SCell}, which should therefore be selected by the mmWave BS to reach the UE with the best performance.

Table 5-1: An example of the complete report table that the MCell builds after having received the partial RTs from the M surrounding mmWave SCells in the considered area. It is supposed that the UE can send the sounding signals through N_{UE} angular directions.

UE direction	SCell_1	SCell_2	...	SCell_M
1	$\text{SINR}_{1,1}$ $d_{1,1}$ $\text{var}_{1,1}$	$\text{SINR}_{1,2}$ $d_{1,2}$ $\text{var}_{1,2}$	...	$\text{SINR}_{1,M}$ $d_{1,M}$ $\text{var}_{1,M}$
2	$\text{SINR}_{2,1}$ $d_{2,1}$ $\text{var}_{2,1}$	$\text{SINR}_{2,2}$ $d_{2,2}$ $\text{var}_{2,1}$	...	$\text{SINR}_{2,M}$ $d_{2,M}$ $\text{var}_{2,M}$
...	...	...	...	...
N_{UE}	$\text{SINR}_{N_{\text{UE}},1}$ $d_{N_{\text{UE}},1}$ $\text{var}_{N_{\text{UE}},1}$	$\text{SINR}_{N_{\text{UE}},2}$ $d_{N_{\text{UE}},2}$ $\text{var}_{N_{\text{UE}},2}$	...	$\text{SINR}_{N_{\text{UE}},N_{\text{UE}}}$ $d_{N_{\text{UE}},N_{\text{UE}}}$ $\text{var}_{N_{\text{UE}},N_{\text{UE}}}$

3) Phase 3: Path switch and scheduling command: If the serving cell needs to be switched, or a secondary cell needs to be added or dropped, the MCell needs to inform both the UE and the cell. Since the UE may not be listening in the direction of the target SCell, the UE may not be able to hear a command from that cell. Moreover, since a common reason for the path switch and for cell additions in the mmWave regime is link failures, the control link to the serving mmWave cell may not be available either. To handle these circumstances, it is proposed that the path switch and scheduling commands be able to be communicated over the legacy 4G cell. Therefore, the MCell notifies the designated mmWave SCell (with ID n_{ID}), via the high capacity backhaul, about the UE's desire to attach to it. It also embeds the best direction d_{SCell} that should be set to reach that user. Moreover, it sends to the UE, through an omnidirectional control signal at microwaves, the best user direction d_{UE} to select, to reach such candidate SCell. By this time, the best SCell-UE beam pair has been determined, therefore the transceiver can directionally communicate in the mmWave band.

Table 5-2: Number of synchronization signals that the BS (or the UE) has to send (and corresponding time) to perform a DL (or UL) based procedure. A comparison among

different BF architectures (analog and fully digital) is performed. We assume T_{sig}=10 us, T_{per}=200 usec (to maintain a overhead of 5%), N_{UE} =8 and N_{BS}=16.

BF Architecture		DL-based SCell transmits UE receives	UL-based SCell receives UE transmits
SCell Side	UE Side		
Analog	Analog	$N_{\mathrm{UE}}N_{\mathrm{BS}}$ (25.6 ms)	$N_{\mathrm{UE}}N_{\mathrm{BS}}$ (25.6 ms)
Analog	Digital	N_{BS} (3.2 ms)	$N_{\mathrm{UE}}N_{\mathrm{BS}}$ (25.6 ms)
Digital	Analog	$N_{\mathrm{UE}}N_{\mathrm{BS}}$ (25.6 ms)	N_{UE} (1.6 ms)
Digital	Digital	N_{BS} (3.2 ms)	N_{UE} (1.6 ms)

5.2.5 Measurements Reporting

In LTE the UE is configured with so-called *mobility events* that contain rules related to the radio measurement of the serving and neighbor cells that are checked by the UE [3GPP16-36331]. For example, the UE sends “RRC Measurement Reports” to its serving cell when the ***RSRP (source) - RSRP (target) > configurable threshold*** for longer than a configurable time-to-trigger. Based on these measurement reports (and other available input), the network may decide to trigger a handover preparation where the source cell requests potential target cell(s) whether a handover may be performed e.g. to check if resources are available. In the case of cells from different eNBs this is done via X2 and S1 procedures. Upon the acceptance from potential target cells the source eNB may send a handover command (via RRC Re-configuration) to the UE and upon the reception of this command the UE should try to access the target cell notified by the handover command (i.e. the UE will trigger a random access procedure followed by a RRC re-configuration). The whole process is summarized in [3GPP15-36300].

As said above, METIS-II assumes that mobility solutions may also rely on measurement reporting (in addition to solutions potentially based on channel reciprocity, mainly in TDD scenarios). An important design question related to that is how often the UE would need to measure the signals and how it should report them. As highlighted earlier a potential aspect that METIS-II is investigating is the fact that in higher frequencies the coverage is likely spottier and more unstable which may affect the timing a mobility procedure should be triggered and how measurements should be reported. This may drive the measurement reporting function to be in another time-scale compared to today’s RRC reporting that occurs in a fashion of few 100’s of milliseconds.

5.2.6 Seamless Mobility

Network controlled handover is typically used to control the mobility of a UE in RRC connected state. There are two types of network controlled handover procedures: break-before-make and make-before-break handover procedures.

In LTE, the legacy handover procedure is a break-before-make procedure. The basic procedure for network controlled break-before-make handover is shown in Figure 5-2. The UE sends a measurement report to the current serving 5G-NB if triggering criterion is fulfilled. The source 5G-NB communicates with the target 5G-NB, e.g. to check if the target 5G-NB has enough resources to admit a new UE. The response to the handover preparation request, i.e. message 3 in Figure 5-2, may include system access configuration, scheduling grants, etc. The source 5G-NB sends a HO command to the UE if the target 5G-NB has sufficient resources to admit the UE. The UE leaves the source 5G-NB and stops transmitting and receiving data after it has processed the HO command. The User Plane (UP) transmission/reception is resumed towards the target 5G-NB after the UE becomes ready for transmission, e.g. after successfully accessing the target cell.

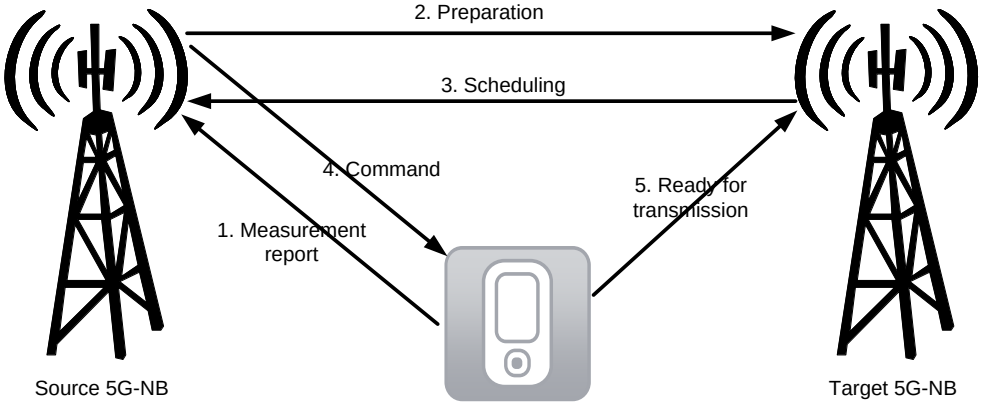


Figure 5-2: Basic procedure for network controlled break-before-make handover

One of the KPIs for handover procedure is interruption time. The handover interruption time is defined as the time duration during which a UE “cannot exchange user plane packets with any base station” during handover [3GPP14-38913].

The handover interruption time in LTE-A is 10.5 ms for FDD RIT and 12.5 ms for TDD RIT [3GPP13-36912] which achieves the ITU interruption time requirement. In 5G, the requirement for interruption time is expected to be tighter than the previous generations [3GPP14-38913]. The interruption time requirement for different applications might be different, some application even requiring an interruption-less handover. Therefore, optimized and diversified handover procedures are required for 5G RAN to meet these requirements.

The interruption time requirement in 5G might not be the same for all use cases. For latency critical use cases it is important to minimize the service interruption due to mobility events for both ideal and non-ideal backhaul scenarios. If the requirement of interruption time is down to 0 ms, then make-before-break mobility events are required implying handovers using dual/multi-

connectivity, for example. Therefore, we propose both break-before-make and make-before-break handover procedures to be considered for 5G RAN. The break-before-make procedure might be a natural consequence of single connectivity; whereas, the make-before-break handover is a natural consequence of multi-connectivity.

Break-before-make handover procedure: challenges and potential solutions

The factors that affect the break-before-make handover interruption time and possible mitigation approaches are discussed as follows:

System information acquisition: The HO command message typically contains configuration of neighbour cells (or nodes) required for system access e.g. PRACH configuration, cell accessibility, etc. Due to this, the UE does not need to acquire the system information of the target cell for initial access. It can start a contention-free RACH procedure without acquiring the system information, e.g. PRACH configuration. Such approach avoids the interruption of packet transmission during handover due to target cell system information acquisition. The approach is also in-line with the envisioned lean system access design for 5G RAN.

Random access procedure: The RACH procedure albeit being contention-free contributes to the packet transmission interruption time during handover. One approach to avoid the packet transmission interruption due to the RACH procedure is through direct data transmission to the target cell without RACH procedure. To achieve this, the 'HO command' message needs to include the System Frame Number (SFN) at which the handover should take place and a scheduling information for the UE in the target cell. This might be challenging in practice. In addition, the UE should be able to estimate the timing advance of the target cell, e.g. as in [BPR+1515], unless the cells are small enough that timing advance is not required. If these requirements can be fulfilled, the UE continues to receive and transmit packets with the source cell until the SFN at which the handover should take place. Afterwards, it can start Uplink (UL) transmission directly to the target cell, e.g. based on the scheduling information received in the HO command. The source cell, however, may need to forward any Downlink (DL) packets that hasn't been transmitted before the SFN at which the handover occurs to the target cell.

Backhaul transmission: The backhaul transmission increases the DL packet transmission interruption time during handover unless the source and target cells are located in the same base station, have the same baseband processor or the backhaul is ideal. Therefore, the DL packet transmission interruption time during break-before-make inter-site handover cannot be reduced less than the backhaul delay.

Multi-connectivity: Make-before-break handover procedure

Multi-connectivity is a natural candidate for enabling make-before-break handover. This may imply UEs with 0 ms interruption time requirement might need to have a capability for multi-connectivity. As per the definition of interruption time, the interruption time during multi-connectivity is 0 ms.

Although the UE can transmit and receive UP packets at any time during multi-connectivity, in practice, there may be delays introduced to individual packets due to mobility during multi-

connectivity sessions. In order to better capture this, additional mobility KPIs might be required. For example, a use-(perhaps packet-) centric packet delay metric might be helpful to capture the delay introduced to individual packets due to mobility procedures during a multi-connectivity session.

One potential approach to minimize the extra delay introduced to individual packets due to mobility procedure is to use multi-connectivity with data duplication. The packets for a user are duplicated and sent across all radio legs. The UE keeps the first packet which is successfully decoded and discards the others. Even with this approach, reducing the extra delay introduced to packets to 0 ms is challenging.

5.2.7 Mobility for Beamforming

Despite the link budget gains provided by beamforming solutions, reliability of a system purely relying on beamforming and operating in higher frequencies might be challenging, since the coverage might be more sensitive to both time and space variations. As a consequence, the SINR of that narrow link can drop much quicker than in the case of LTE.

Using antenna arrays at access nodes with the number of elements in the hundreds, fairly regular grid-of-beams coverage patterns with tens or hundreds of candidate beams per node may be created. The coverage area of an individual beam from such array may be small, down to the order of some tens of meters in width. As a consequence, channel quality degradation outside the current serving beam area probably due to (small) objects that shadow the beam, is quicker than in the case of wide area coverage.

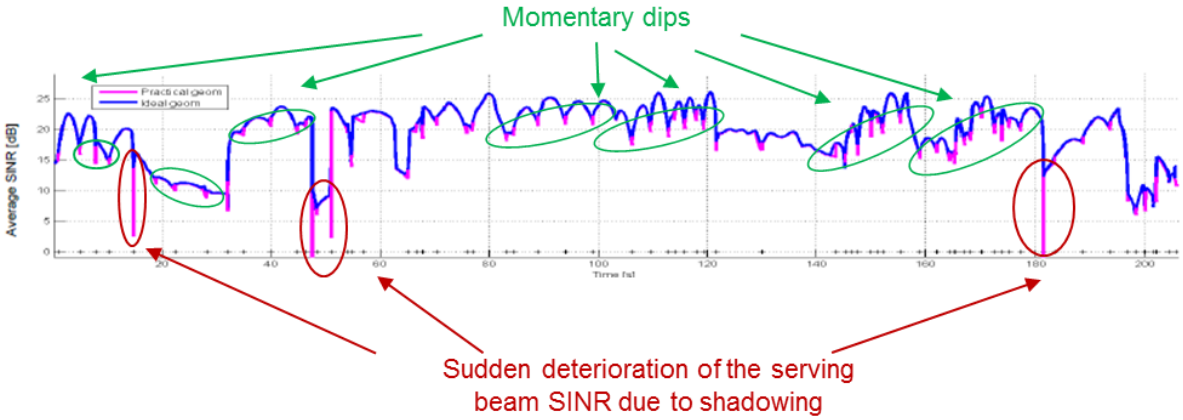


Figure 5-3 Quick SINR drops due to shadowing of the beam to the source

Figure 5-3 shows an example of average SINR variations for one UE and one single cell employing a large-array grid of beams in a 15 GHz carrier frequency. An optimal beam selection scheme is compared with a simplified measurement-based scheme based on an estimated SINR and reference signals transmitted over 6 PRBs and 1 subframe (equivalent to 144 samples). The comparison of the simplified measurement-based scheme to the optimal serving beam choice shows some momentary dips below the optimal performance (It has also been observed that the

SINR distributions are hardly affected by these dips at most vehicular speeds). On the other hand, we see also some deeper dips that indicate a sudden deterioration of the serving beam SINR due to shadowing, e.g. in “around the corner” situations. The serving beam SIR may drop by over 20 dB within 5-10 ms. Further results have also shown that these occasional drops will be unavoidable at 10-30 GHz. One way to overcome this is to use a beam-centric mobility and multi-connectivity, see Figure 5-4.

Figure 5-4 Beam-centric mobility and multi-connectivity

Infrequent interruptions due to low SINR occurrences may not be critical for the User Plane (UP) flows. However, Control Plane (CP) flows transmitted over these beams might be carrying mobility related signaling such as transmission of measurement reports and reception of handover commands which might lead to many error cases, leading to high handover failures and radio link problems. Thus, the overall aim of the active state mobility is to maintain the connectivity between UE and network during UE movement and radio environment change. However, different mechanisms might be required to fulfill this goal depending on the association of the beams to the logical/physical network architecture. This is demonstrated by the following example in Figure 5-5:

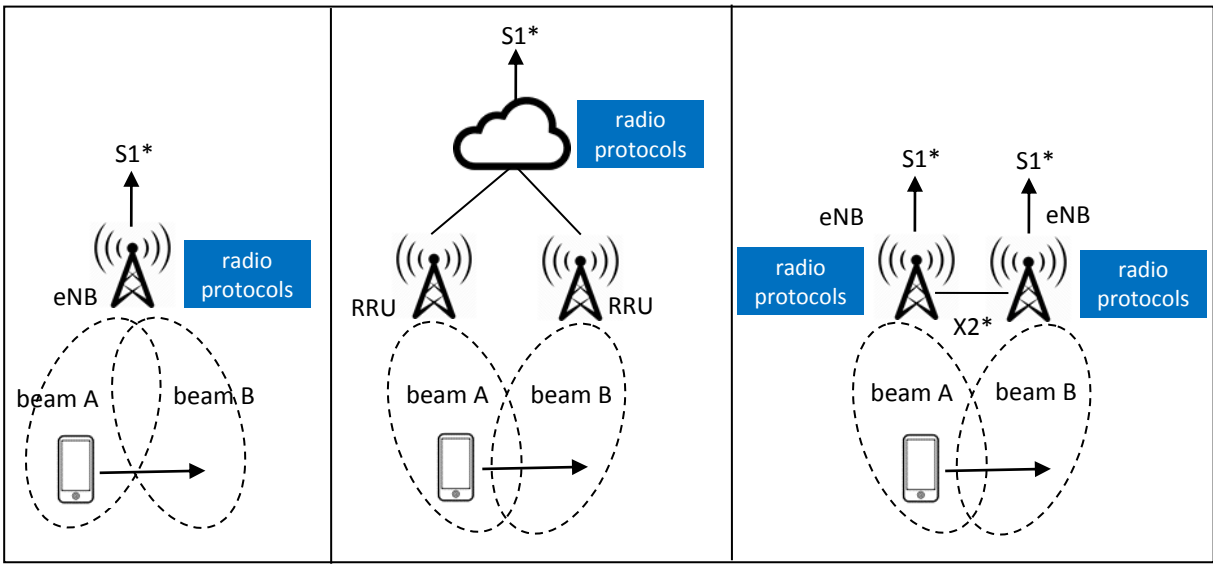


Figure 5-5: Scenarios for beam change

1) Beams A and B are transmitted from the same network node, which hosts common radio protocol instances (PHY, MAC, RLC, PDCP, RRC) and common core network interface (S1*) for the two beams.
2) Beams A and B are transmitted from two remote radio units (RRU) which are connected to a common cloud node. The interface between RRU and cloud node is ideal.
3) Beams A and B are transmitted from distributed network nodes that host separate radio protocol instances and separate core network interfaces for the two beams. The interface (X2*) between network nodes is non-ideal.

It can be observed that scenarios 1 and 2 provide some additional degrees of freedom over scenario 3 in terms of beam management:

- Beams A and B are managed by the same protocol instances → fast/centralized coordination of the beam switch, no protocol reset.
- Beams A and B have a common core network termination point → no CN path switch.

In order to exploit these benefits, each beam needs to be associated with a network location identifier (cell). However, the use cases for such cell identifier are not necessarily limited to the ones of legacy systems such as LTE.

The main characteristics of intra-cell (or intra-node) and inter-cell (inter-node) mobility are described in the following.

Intra-cell mobility

Intra-cell mobility refers to a scenario where the UE is communicating with the network using beams with the same cell/node identifier. The beams of one cell may be transmitted from a single geographical transmission point or from multiple points in a coordinated manner.

The key assumption is that the logical network node terminating the radio interface protocols remains the same when the UE moves to a new beam. Hence, the radio interface protocols may continue without change or re-establishment and no packet forwarding is needed. The cost of beam ping-pong in terms of inter-node signaling is small and comparable to precoder selection of LTE.

The requirements for intra-cell beam mobility can be then summarized as:

- The design should support a fast switching/tracking of the communication beam to combat rapid changes in link quality.
- The design should be able to exploit the availability of multiple overlapping beams that can be used for the communication with a single UE.
- The design should be able to exploit co-location of the radio protocols for the source and target beam.
- Beam mobility should have a minimum impact to the RRC layer.

The detailed design is being studied in METIS-II, addressing e.g. the following aspects:

- What is the optimal protocol layer for the intra-cell beam management?
- What are the required reference signals for UE and/or network to support beam management?
- What are the mechanisms for UE based feedback? Potential alternatives include periodic, network requested, and trigger based reporting mechanisms.
- What is the required time scale or time scales of the reporting mechanisms?
- Is the beam management controlled by UE, network, or both?

Inter-cell mobility

Intra-cell mobility refers to a scenario where the UE switches its communication path from a set of beams sharing one cell identifier to a set of beams sharing another cell identifier. Such cell change resembles the inter-site handover of LTE, which typically comprises a neighbor cell measurement report from the UE, handover preparation with inter-node signaling, and handover execution with data forwarding between source and target nodes.

However, the possibility for a rapid degradation of the radio link (up to 20 dB in 5-10 ms) as described above and increased handover rate (up to one handover per second) [TCG+14] introduces challenges for the high-frequency inter-cell mobility that are not present in the legacy low-frequency systems.

In LTE, the total time for the UE to send a measurement report and receive a handover command from the source cell is shown to be in the order of 40 ms [BPR+15]. This is clearly not fast enough to combat a fast channel degradation of a time scale of 5-10 ms.

The increased handover rate implies to an increased signaling load on RAN and possibly CN interfaces.

The key requirements for inter-cell mobility can be then summarized as:

- Fast HO preparation and execution to combat rapid channel degradation.
- Avoid path switch to reduce the impact of increased handover rate.

One potential technique to meet these requirements, studied by METIS-2, is the cluster based mobility. A cluster is a set of cells that the UE can detect and which are prepared in advance for a fast re-routing of the signaling and user data. The basic principles of this scheme are illustrated below:

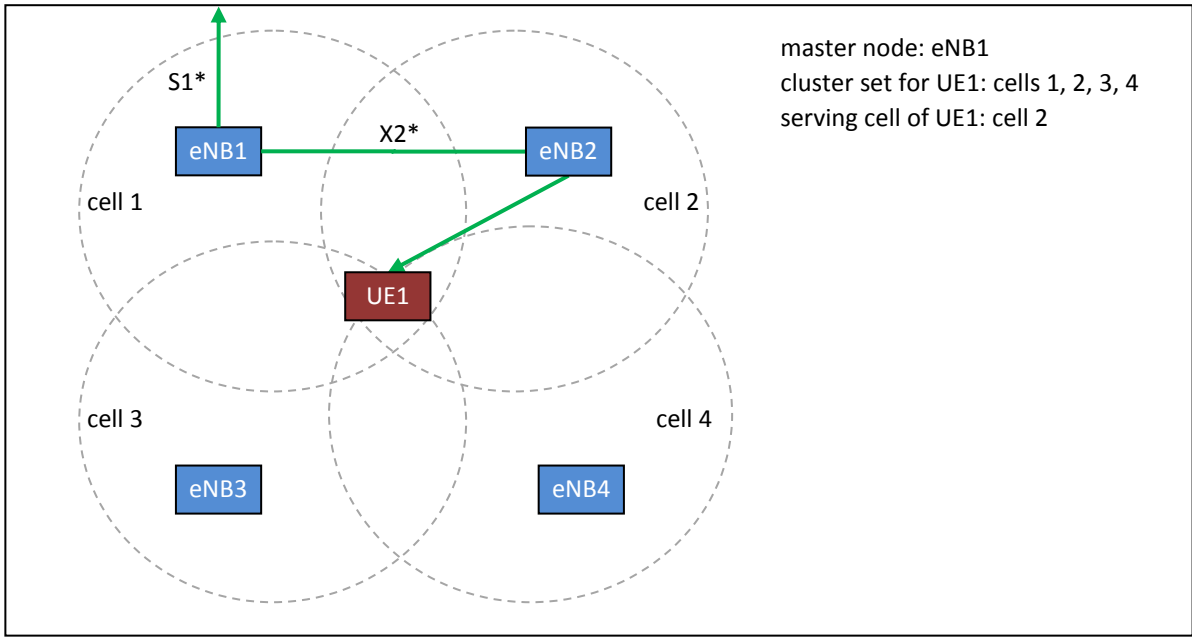


Figure 5-6: Cluster based mobility

A cluster set is managed by a master node. The master node hosts the RRC connection, terminates the core network interface, and has the responsibility of establishing the data and control planes of a UE for a given time. The other nodes in a cluster set, called slave nodes, are candidates for fast re-routing upon link blockage or degradation.

A link degradation is detected by the serving cell based on some fast mechanism. The network node hosting serving cell then sends a re-routing request to the master node. Master node then selects a new serving cell based on a priority list that accounts e.g. the loads and signal qualities of the cells in the cluster set. Finally, the master node switches the UE path and starts forwarding data to the new cell.

The UE measurements are utilized in cluster based mobility for helping the master node to add or remove cells from the cluster set and for forming the priority list of cells for the fast re-routing.

5.2.8 High Speed Design

In high speed scenarios e.g. 240 Km/h or even higher up to 500 Km/h, channel estimation is quite difficult and the inter-carrier interference is significant due to large normalized Doppler frequency. Conventional Doppler effect cancellation methods in the receiver in physical layer can possibly be used to improve system performance. However, these methods have heavy computational complexity or large redundancy, and the performance will also be restricted by inaccurate channel estimation and large normalized Doppler frequency. Therefore, the considerations of increasing reference signal density and increasing subcarrier spacing for high or very high mobility scenarios can be used to effectively alleviate the above issues and support high or very high mobility requirements without heavy computation complexity or large redundancy in the receiver.

When a UE moves slow, the Doppler effect is small, i.e. inter-carrier interference can be neglected. However, when the UE moves fast, the ICI will be significant. After neglecting noise term, the OFDM received signal with ICI can be expressed as:

$$Y_k = X_k H_{k,k} + \sum_{\substack{m=0 \\ (m \neq k)}}^{N-1} X_m H_{k,m} = X_k H_{k,k} + ICI_k$$

X_k, Y_k, ICI_k are the frequency domain transmitted signal, received signal and ICI terms at k-th sub-carrier, respectively. Besides, $H_{k,m}$ is the channel frequency response from m-th sub-carrier to k-th sub-carrier.

Normalized Doppler frequency δ defined as $f_d * T_U$ where f_d is the maximum Doppler frequency and T_U is the useful symbol duration is generally used to indicate the influence of the Doppler effect. When the normalized Doppler frequency is large, the ICI will also be large and then the system performance will be degraded significantly. Besides, channel estimation will be another system impact due to Doppler effect because the channel is fast time-variant and will be not easy to estimate correctly when the normalized Doppler frequency is large.

The high speed scenario is suitable for the use case of connected cars that will support 140 km/h for motorway scenario, i.e. the maximum relative vehicle speed will be up to 280 km/h or even higher. Although the carrier frequency currently in rural macro (RMa) deployment is 800 MHz in the motorway scenario of UC5 connected cars, the carrier frequency above 6GHz should be considered in the future requirements. Besides, one of the 5G requirements is to support a relative speed of up to 500 km/h. Therefore, the Doppler effect is a significant issue in the high speed scenario and RAN design for the high speed scenario should be considered to support high or very high mobility requirements without heavy computational complexity or large redundancy in the receiver.

Increasing DMRS density to 4 symbols per 1 ms with reusing PUSCH DMRS sequence in each physical sidelink channel except for PSBCH is the working assumption in TR 36.885 [3GPP16-36885] for handling high Doppler case. However, increasing DMRS density is not enough for high

order modulation, for example 16 QAM, in the high Doppler case if advanced ICI cancellation and channel estimation methods are not adopted in the receiver. The details are shown in Annex A.4.

When we increase subcarrier spacing, the normalized Doppler frequency will be decreased and hence the Doppler effect will be mitigated without any advanced ICI cancellation design in the receiver. Then, the ICI will be mitigated and the system performance will be improved without any advanced ICI cancellation design in the receiver. Therefore, increasing subcarrier spacing, i.e. reducing symbol duration in the subframe structure, is a good RAN design consideration to support high or very high Mobility requirements without heavy computation complexity or large redundancy in the receiver. For example, UE can report the geo-location to RAN (eNB). Based on this received information from UE, RAN determines which type of subcarrier spacing (e.g. low speed type, medium speed type, and high speed type scenarios) is suitable for the particular link. Then, RAN informs the UE pair the particular type of subcarrier spacing they should use. An Illustration of increasing subcarrier spacing (for example, subcarrier spacing increased to 30 KHz) from legacy LTE (subcarrier spacing is 15 KHz) is shown in Figure 5-7.

However, in order to achieve the high data throughput requirements in high speed scenario, high order modulation, for example 16 QAM modulation, should be used in the data channel. And hence, the system performance in high order modulation will be more sensitive to both inaccurate channel estimation and ICI in the receiver. Increasing subcarrier spacing to mitigate ICI may not be enough because the system performance is more sensitive to inaccurate channel estimation in the high order modulation case.

Therefore, in order to consider both channel estimation and ICI issues in the high speed scenario, flexible adjustment or selection for increasing subcarrier spacing and increasing reference signal density according to the modulation types and normalized Doppler frequency, i.e. carrier frequency and maximum relative speed, is a suitable approach for the RAN design considerations that is proposed to support high or very high mobility requirements without heavy computation complexity or large redundancy in the receiver.

An llustration of the proposed method (for example, subcarrier spacing is 30 kHz and 3 DMRS symbols per subframe) from legacy LTE (subcarrier spacing is 15 kHz and 2 DMRS symbols per subframe) is shown in Figure 5-8, and performance comparisons among legacy LTE, increasing DMRS density (subcarrier spacing is 15 kHz and 4 DMRS symbols per subframe), increasing subcarrier spacing (subcarrier spacing is 30 kHz and 2 DMRS symbols per subframe) and the proposed method (subcarrier spacing is 30 kHz and 3 DMRS symbols per subframe) for 16 QAM in the high Doppler case for V2V is shown in Figure 5-9. From the simulation results, we observe that the proposed method, i.e. increasing subcarrier spacing to 30 kHz and increasing DMRS density to 3 DMRS symbol per subframe, still performs well even in 16QAM modulation in high Doppler case.

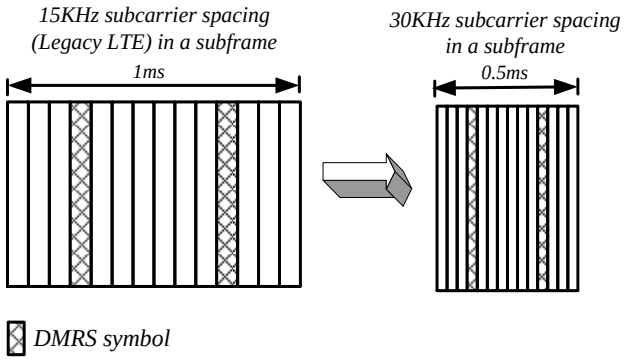


Figure 5-7: An Illustration of increasing subcarrier spacing from legacy LTE

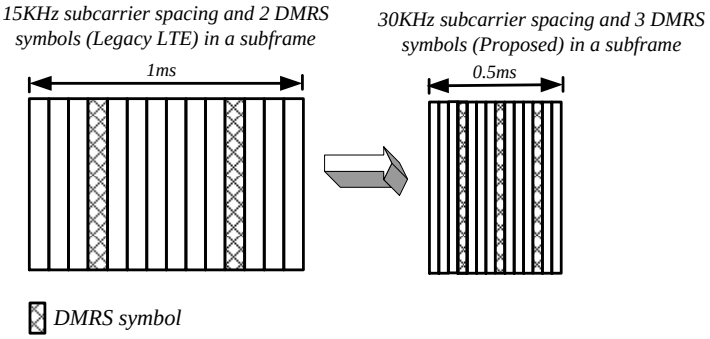


Figure 5-8: An Illustration of the proposed method from legacy LTE

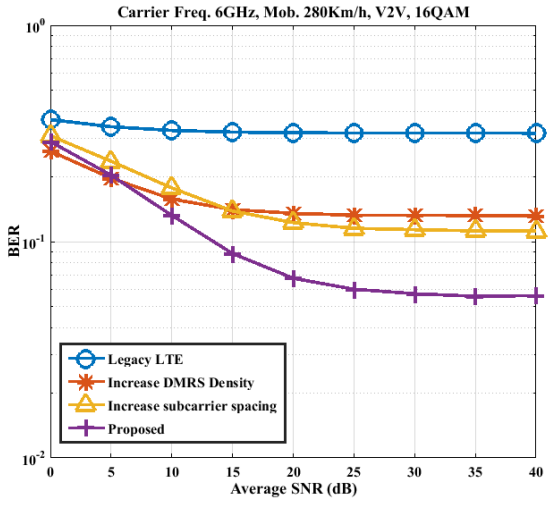


Figure 5-9: Performance comparisons among legacy LTE, increasing DMRS density, increasing subcarrier spacing and the proposed method for 16QAM in the high Doppler case

5.3 Connected Inactive Mode Mobility

5.3.1 Introduction

This section gives a high-level description of mobility management of always-connected UEs during the new proposed low activity mode mobility called “RRC Connected Inactive". In this section, we describe potential ways on how to operate the state transitions between RAN and devices and what new functionality to expect for 5G UEs and network.

5.3.2 RRC Connected Inactive state Management

Tracking Area Management

In LTE, the tracking area (TA) management to locate the UEs is centralized to MME (Mobility Management Entity). To allow always connected UEs in 5G, the location tracking could be done in RAN since the mobility anchor and connection to core network is available. Distributed TA management function in RAN has information available about their adjacent TAs in order to create and deliver up-to-date “allowed TA list” to the UEs that are configured to RRC Connected Inactive state.

According to Figure 5-10, the 5G-NBs are RAN nodes, which in this example serve four 5G radio areas. For example, 5G-NB1 is serving 5G areas 11, 12, 13 and 14 in the Figure 5-10. Each 5G-NB connects to a control-plane node in the core network e.g. mobility management by using the enhanced S1*-c interface and to the User Gateway by using the enhanced S1*-u interface. The neighbouring 5G-NBs are inter-connected by using enhanced X2* interface at least on the control-plane.

When a UE is in Connected Inactive state, it receives from its serving 5G-NB one or more list of cells belonging to allowed TA. According to Figure 5-10, the UE gets configuration with Tracking Area 1 and 2 and UE can move within cells in these TAs during Connected Inactive state without informing the network about the cell reselection during the low activity periods. The serving 5G-NB does not need to perform the serving 5G-NB change even if the UE performs cell reselection and informs network about the new Cell Id.

Each 5G cell advertises its Cell Id and Tracking Area ID(s) in its broadcast channel. The cell changes to different 5G-NBs in the allowed TAs are handled transparently to the UE with assistance of the current 5G-NB that takes care of reporting UEs current location to the UEs’ Last Serving 5G-NB.

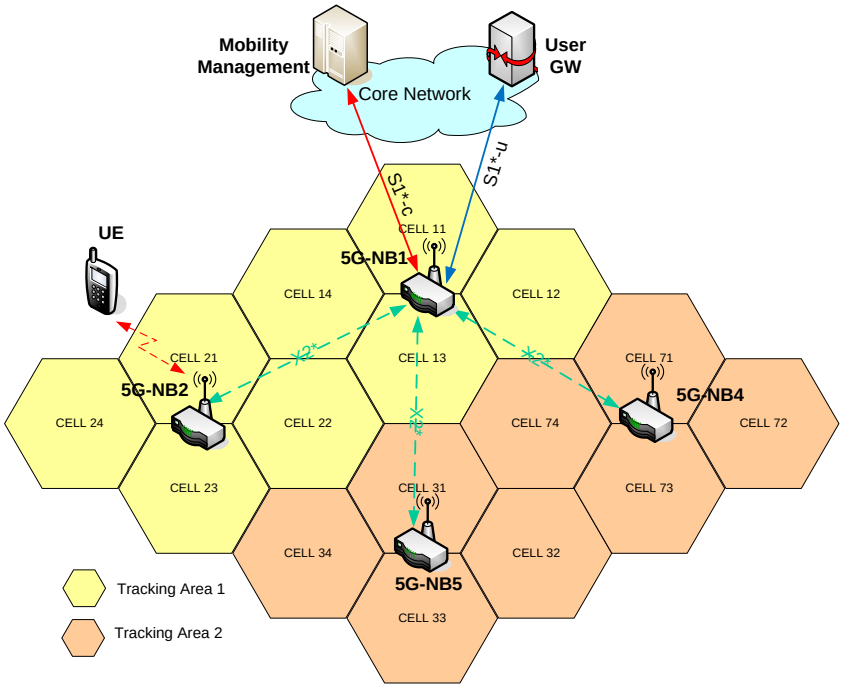


Figure 5-10, 5G Access Network and tracking areas

Location tracking during RRC Connected Inactive state

Mobility during Connected Inactive state may cause frequent path switching. Therefore, the UE context transfers due to node reselections can be reduced by keeping the UE context and the data path(s) terminated in the Last Serving 5G-NB where UE was in RRC Connected state. Now the Last Serving 5G-NB can take the role of a mobility anchor, which allows keeping the C-plane and the U-plane RAN connections unmodified towards the 5G Core Network.

The 5G-NBs in which UE is located during Connected Inactive state registers the UE's current TA/5G-NB (i.e. location) to the Last Serving 5G-NB by indicating its Last Serving 5G-NB ID to the 5G-NBs it is visiting during Connected Inactive state.

The Last Serving 5G-NB shall initiate the paging when receiving downlink packets and deliver the UE context data and the buffered DL packets to the current 5G-NB UE is visiting during Connected Inactive state. The 5G-NB in which UE is visiting shall become the new serving 5G-NB that shall re-configure UE to connected state and performs the required path switching procedure.

In the UE initiated connection the current visited 5G-NB will retrieve UE context from the Last Serving 5G-NB based on UE reported Last Serving Cell Id and buffer UL packet(s) until it has performed the required path switching. In case the UE has not moved from its Last Serving 5G-NB, then the UE context is instantly available and there is no need to perform any path switching.

In the distributed TA management, the Last Serving 5G-NB may detect UE movement out of the current TA based on its received UE registrations. Last Serving 5G-NB should deliver UE AP

context to the new 5G-NB to let it to take the role of Last Serving 5G-NB, and to trigger the TA Update procedure and to perform path switching.

It is worthwhile to note some use cases, e.g. low latency applications, which might require switching the S1* connection to the optimal 5G-NB location as soon as possible. For example, the S1* connection might need to be immediately switched when the UE moves to a new cell in an area which is not located in the 5G-NB currently terminating the S1* connection. Further, the network can consider the deployment aspects and for example resume the connection from macro cell in the same TA.

5.3.3 Mobility Benefits of RRC Connected Inactive in 5G

Mobility in LTE is network controlled during the RRC_CONNECTED and UE controlled during RRC_IDLE state. According to proposed state model in Figure 3-3, the UE can take mobility control when the RRC state is RRC Connected Inactive. Some of the mobility benefits of the proposed state model are:

- Provides a state model where the CN/RAN context is stored allowing forwarding of incoming packets from the CN to the latest mobility anchor point at the RAN.
- Distributed Tracking Area management in RAN allows UE centric or network slice specific TA optimization depending in the user mobility. Stationary UEs can have TA defined as a single cell minimizing the signaling related to paging or even making paging obsolete.
- Network can benefit from reduced signaling towards a core network when a connected UE selects a new cell during low activity period.
- Last serving cell can control S1* relocation and/or path switch when a UE needs user-plane connectivity to send data. In some cases it is useful not to relocate the S1*, e.g. when the UE is a stationary MTC device generating small data using contention based uplink transmission.
- Mobility procedures are simplified with lower signaling overhead when the UE can transmit contention based uplink data while still in the RRC Connected Inactive state.
- State can support highly configurable procedures for contradictory requirements of various 5G use cases in terms of reliability, mobility, latency, bandwidth, battery life, etc.

5.4 Context Aware Mobility Management and Traffic Engineering

5.4.1 Introduction

This section describes methods to improve the mobility functionality using context aware mechanisms. Since all the information used to enhance the decision making process of an entity is used context, the context aware mechanisms include a huge variety of mechanisms. In general the context aware mechanisms that improve mobility are mainly based on real time information,

thus they tend to be inefficient and lack accuracy since they are based only on the real time measurements and do not consider the past user behaviors.

In this section two schemes are presented focusing on behaviour prediction using data analytics and predicting the user mobility using Markov chains. The former, by analyzing the past user behaviour extracts multi-dimensional behavioural profiles which are stored for future use. The context information is gathered and processed offline, thus reducing the complexity and signaling cost of the scheme, whereas the extracted profiles are being combined with online information so as to predict more accurately the user behavior. The latter moving towards similar direction exploits past direction knowledge so as to optimize the UE trajectory prediction by considering also the UE origin and destination; the UE mobility is modeled using a Markov chains.

5.4.2 Data Analytics for Traffic Engineering

According to the NGMN white paper [NGM15], the end users will have access to a diverse set of services (ranging from high definition video and audio, web browsing, games, to keep alive messages etc.). For the provided services (e.g., augmented reality, cloud services, car to car communication, etc.) connectivity everywhere (including indoor and outdoor environments, environments with ultra-dense or limited infrastructure, or where the environment is extremely crowded, etc.) to users with various mobility shall be provided [NGM15]. Thus, the network needs to meet at the same time the data rate demanding services (such as high definition video and audio) and the low demanding ones (e.g., keep alive messages) which may however involve millions of devices accessing simultaneously the network; all the user equipment's may have diverse mobilities.

In such demanding environments the use of context information is imperative for enabling the efficient use of the network resources [MSC13]. Also according to [MSC13] the term context refers to *"all the information that may be exchanged among the network entities, even heterogeneous ones so as to solve challenging networking problems such as management and control of the network resources"*. This generic definition of the context information implies that all information types may be used for the optimization of the network management and control such as radio measurements, mobility information/measurements (e.g., user speed, direction, etc.), etc.

Several Context Aware mechanisms are available in the literature. Such schemes take advantage of e.g.,

- Location information [XPM+14] [CWH+13],
- Mobility information [KKP+13],
- Network statistics (access or backhaul) [QXY+10] [MBS+10],
- Current Service information [XPM+14] [MCC+11].

Combination of the abovementioned parameters can improve the accuracy of the mechanisms. In general such combinations are based on objective functions and/or rule based systems, which

however try to combine inputs not clearly linked to each other. The solution presented in [TP10] is an example of such a scheme.

However:

- In these context-aware systems the decisions require real time information related to a significant number of parameters
- Even enhanced with state optimization, the rule based systems cannot address the "Curse of Dimensionality" ([SRG+12]) efficiently and in real time [MKS+10]
- Transferring all these inputs to serve every radio resource management and control mechanism for all users in an area is costly for processing (e.g., hundreds of requests per second)
- Collecting all necessary information in real time may also introduce unnecessary delays for delay critical functions (e.g., interference schemes).

This information for being transferred among the network elements is very demanding and requires the consumption of huge amount of resources for the transfer and the handling/processing. Real-time context aware mechanisms may require a considerable rate of transfer exchange. According to [PF13] up to 2,618 Connections per device and per day may be assumed thus leading for special use cases to a huge number of interactions with the network per second. Figure 5-11 shows, using a simple analogy, the number of connection requests per second using METIS I ultra-dense networking use cases as reference.

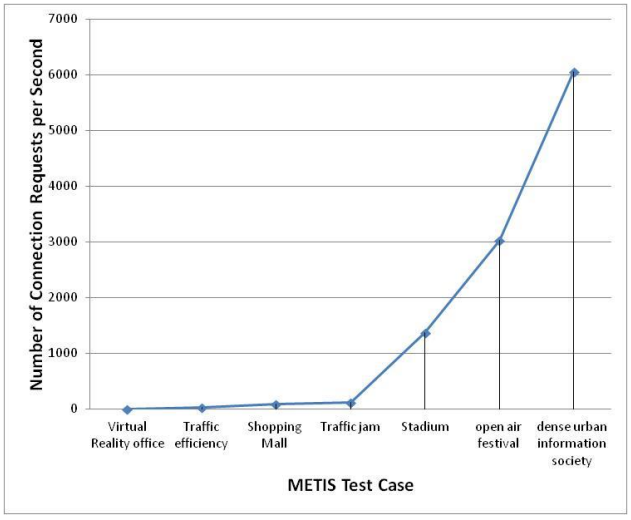


Figure 5-11: Number of connection requests per second per area for METIS I ultra-dense networking use cases.

Context aware mechanisms require real time access to several information elements such as location and positioning, speed, throughput, RSS, SINR, for access or backhaul links, service information, user Profile/Preferences, etc. [MSC13]. The combination of the previous information

elements enables more accurate decision making. However, such combination is a demanding task in terms of:

- Accessing, processing, and combining in real time many information elements
- Identifying the user preferences in relation to specific parameters such as, location, date, time, user equipment capabilities, battery level, charging level, educational level, etc.

Combining these information elements increases the number of variables, which consequently increases the problem state space, a problem known in the research area as the “curse of dimensionality” ([SURDIM]). This term refers to the fact that in the absence of simplifying assumptions the sample size needed to estimate a function of several variables to a given degree of accuracy in order to get a reasonably low variance estimate grows exponentially with the number of variables.

Regarding the user preferences and behavioural profile used in context aware mechanisms, 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 behavioural 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).

Specifically, the users tend to change their behaviours 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 so as to better manage the network resources.

A simple example of modelling the user behaviour 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.

Context aware mechanisms for predicting the future behaviour of users, exploit current information (i.e., current contextual information) so as to identify the user behaviour (e.g., user mobility, the service that the user will access and/or the duration of the service access, etc.) so they may have accuracy problems since they do not use past knowledge, or they cannot make predictions for longer time periods. On the other hand, using past context information to identify user behavioural models will enable more accurate predictions.

By using data mining mechanisms, we may identify the most impacting parameters in the user behaviour. Afterwards learning mechanisms can build user behavioural 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.

Thus,the analysis of the past user actions enables the extraction of multi-dimensional behavioural profiles for future use. The dimensions of the profile could be location/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 behavioural profiles. The profiles are being distributed to the network components and used in online manner by combining online information with the profiles for more accurate prediction. The predicted behaviour is then used for the network operation. Afterwards, the overall network performance has to be evaluated and fine-tuned according to the effectiveness in the network performance.

Specifically the process will consist of the following phases:

- The offline phase :
    - Gathering the user personal context and history and store it in a logically centralized entity
    - Use of this collected information for identifying the most relevant parameters.
    - Extraction of the user behavioural profiles that capture the predicted services, per several variable parameters such as to location, time, subscription, battery level, charging status, mobility pattern, etc.
    - Formation of a multi-dimensional profile for each user for each one of the previous parameters.
    - Storing the multi-dimensional profiles.
- The online phase:
    - The distribution of the profiles to the networking entities. The networking entities include but are not limited to user equipments, Access Stratum and Non Access Stratum control entities, databases etc.
    - The integration of the profiles in the network management and the combination with online information, so as to enable more accurate prediction of the user behaviour.

This behavioral prediction enables the prediction of the overall needs of a group of users in a certain area with high granularity. Specifically, knowing the user behavior with high accuracy will enable the proper placement of the users on the one hand and the prediction of the overall traffic requirements. This will further facilitate the network operator to proceed in the respective actions so as 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). The process of storing and distributing the profiles, as well as the requirements prediction has a direct impact in the RAN. In particular, the RAN should be updated in such way so as to describe with higher granularity the user predicted behavior. Up to now, this 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, mobility. This index however is able to describe only a 256 behaviors and lacks flexibility thus making the description of the user behavior very static. With the multi-dimensional profiling the user behavior will be accurate and the network actions more targeted.

5.4.3 Diurnal Mobility Prediction to Assist Context Aware RRM

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. The mobility prediction accuracy can be enhanced by using additional context information of user origin, destination. The mobility prediction can then be used in tandem with resource allocation to enable services like streaming media to be sustained even in coverage holes.

To enable such mobility context awareness, additional context information, appropriate signaling and interfaces are required. These details along with algorithm are discussed in this technology component

There are several works in literature focusing on mobility prediction such as, prediction based on distributed markov chains, hidden markov model based prediction [SWY+10], making use of neural networks and machine learning [LH05] etc. A majority of these works consider regular hexagonal cells and intend to predict the next cell for user transition based on different strategies. Further, they consider either straight line mobility of users or random way point mobility, to evaluate the performance of prediction schemes.

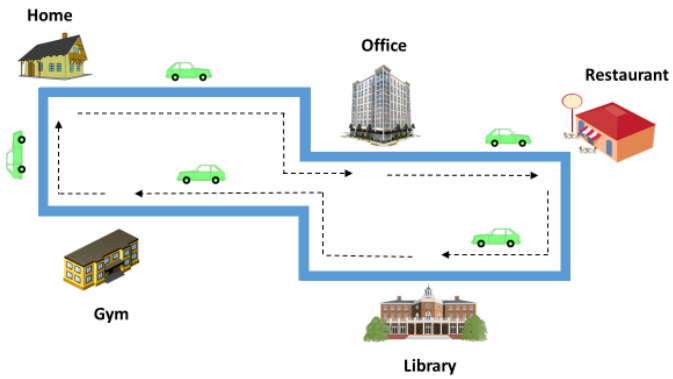


Figure 5-12: Diurnal mobility example

In real world scenarios, the user mobility is not random but is rather direction oriented, see Figure 5-12. The user direction relies on its origin and destination. Further, there are several users who exhibit similar mobility patterns on daily basis. They tend to regularly traverse a limited set of trajectories, comprising of specific landmarks. For instance, an office goer or commuter in public transport takes similar trajectories on regular basis. 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 (next cell prediction, future route prediction).

In simple Markov based prediction, at each cell statistics of number of times a user transited from $cell_n$ to $cell_{n+1}$ is considered. Based on mobility statistics of the user over several business days, probability of transition into a next cell is obtained using Markov chain as,

$$P(BS_n \longrightarrow BS_{n+1}) = \frac{N(BS_n \longrightarrow BS_{n+1})}{N(BS_n)}$$

Where $BS_n \longrightarrow BS_{n+1}$ indicates transition of user from cell n to cell n+1, $N(BS_n \longrightarrow BS_{n+1})$ indicates number of times a user in cell n transited to cell n+1, $N(BS_n)$ indicates number of times user was found in cell n.

Markov based prediction can be extended by using information about origin of the user. In this case, the statistics of number of times a user transited from $cell_n$ to $cell_{n+1}$ is considered provided a user started from a specific origin. The probability of transition is obtained as,

$$P(BS_n \longrightarrow BS_{n+1}/Origin) = \frac{N(BS_n \longrightarrow BS_{n+1}/Origin)}{N(BS_n/Origin)}$$

The model can be further extended using information of both origin and destination. In this case, the statistics of number of times a user transited from $cell_n$ to $cell_{n+1}$ is considered provided a user started from a specific origin and is travelling to a given destination. Probability P of transition is obtained as,

$$P(BS_n \longrightarrow BS_{n+1}/Origin\&Destination) = \frac{N(BS_n \longrightarrow BS_{n+1}/Origin\&Destination)}{N(BS_n/Origin\&Destination)}$$

Similar to the above predictions, route prediction could also be done. More details and elaborated examples could be found in Annex A.5.

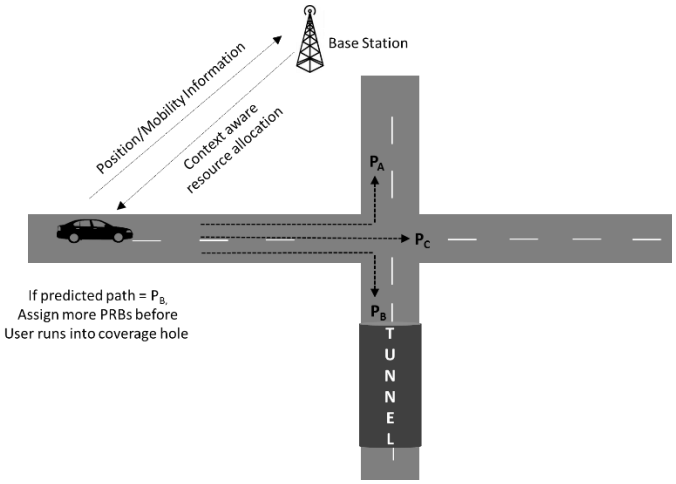


Figure 5-13: Context aware resource management

The mobility prediction concept can be used to design context aware RRM during practical situations like coverage holes. By predicting the movement of user into a route with coverage hole in near future, smart resource allocation can be triggered to support certain services even in coverage holes. The basic idea of the scheme is shown in Figure 5-13.

The above mentioned concept is implemented in madrid grid [MET13-D61] with 7 landmarks at different sites. A set of 10 trajectories are used with each trajectory including all landmarks, constituting diurnal mobility. A 100 simulation runs are executed to obtain mobility statistics and another 100 simulation runs are used to do prediction and average prediction accuracies are obtained as in Figure 5-14. The additional context information of origin and destination is shown to enhance the accuracy of mobility prediction.

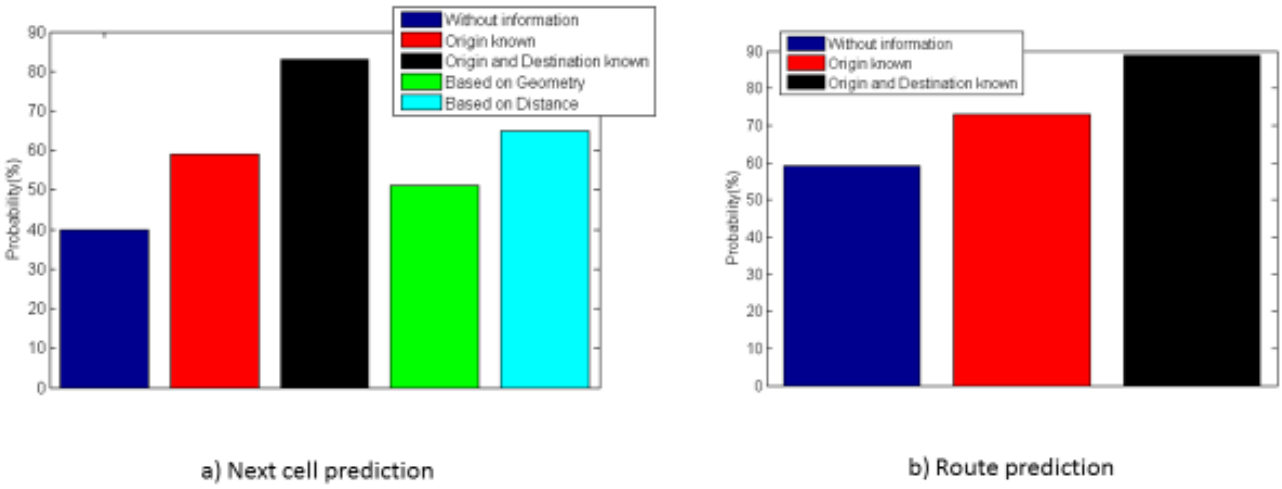


Figure 5-14: Mobility Prediction Results

More elaborations on prediction schemes and context aware resource allocation could be found in Annex A.5. Further, overall architecture with required signaling and interfaces could be found in Annex A.5.

5.4.4 Benefits of Context Aware Mobility Management and Traffic Engineering

Context aware networking, since it considers apart from typical radio measurements such as RSS/RSRP, RSRQ, etc, a big number of inputs, may significantly optimize the decision making process. By considering inputs such as the time, the mobility direction, the destination and the starting point, the network may predict with high certainty the user mobility pattern, as shown in the previous subsection. This enables further the network to make proactive decisions for e.g., reserving resources on the next cell based on the prediction where the user will be in the future.

The use of data mining enables further, on the one hand the extraction of the user preferences and on the other the accurate prediction of the user behavior, since this based on his previous decisions. The nature of the data mining mechanisms, requiring a rather large dataset to be used as training set, implies that the data have to be collected and processed offline, which further reduces the signaling and computational overhead. This way the network may extract the user preferences automatically which will be reflected in the users' behaviors. By combining the overall user preferences captured in multi-dimensional profiles with online information (e.g., time, date, location, UE battery level, charging status, etc.) the network may make very accurate predictions of the user preferences, since they will be linked to the previous user actions. This behavioral profiling enables optimal user placement in the cells/layers and overally optimal traffic engineering.

6 Native D2D Support

6.1 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. 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 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. Thus requiring several considerations and also imposing challenges from CP design perspective e.g. to support efficient and optimal resource management and channel access, unified addressing, cooperative communication, network offloading and inter-RAT/intra-RAT Mobility Management. Below subsections provide details of these and other related aspects of D2D in 5G.

Among others, one important D2D communication scenario and architecture considered here involves D2D Relays and wireless self-backhauling. Self-backhauling refers to the case where a device or network node appears and acts as an access node towards other devices, but uses the same cellular technology to establish a wireless backhaul link towards an access point. In 5G, this is seen as especially important in the context of either very dense deployments, where it may not be economically feasible that all access nodes have wired backhaul available, or in cases where ordinary devices may act as self-backhauled access nodes to provide coverage extension to other devices.

The concept of D2D has been there since quite some time e.g. WiFi Direct, Bluetooth, etc. Also, 3GPP has developed specifications related to D2D, 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 is specified in [3GPP15-23303]. 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 [3GPP15-36211]. 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 SIB or RRCConnectionReconfiguration message and Rx pool is provided in 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 [3GPP16-36331], 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 [3GPP16-36311]. 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. Several of these and other challenges and opportunities are discussed in below sub-sections.

6.2 D2D Enabled and Group based RACH Access

Random access of a large number of devices will introduce new challenges to the 5G RAN design. Thus a new approach to address this problem in 5G is required. Here we discuss a potential solution especially for a group of devices that are static or semi-static. Devices in the same group uses unicast D2D communication or one-to-many/all D2D communication. Besides, a device, based on various criteria depending on particular deployment scenario, is selected as a group head or cluster head. Group head uses PC5* interface towards its directly connected devices for D2D discovery and/or communication. Group head uses Uu interface for communication with the 5G-RAN. In this case, instead of having all the group members to proceed in random access using one of the 64 preambles when they have to transmit, the transmission requests could be aggregated, and only one device (i.e. the group head) will perform the random access request. Thus instead of having all the devices competing for resources, only the group heads will compete. The technique has the potential to reduce 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. Group heads could further coordinate among each other by intra cluster communication i.e. D2D unicast communication or one-to-many communication among group heads, as depicted in Figure 6-1. Below are the main aspects of this scheme:

a. The devices are being grouped by the network based on their mobility and communication characteristics (e.g., data to be transmitted, packet delay requirements)

b. The network schedules the cluster heads' transmission opportunities based on their transmission requirements. The scheduling information includes how many timeslots each device should attempt to access the network and also which preambles should be used.

c. The intra cluster communication may take place either via a different interface e.g. IEEE 802.15/Zigbee or IEEE 802.11 or via D2D communication over 5G AIV.

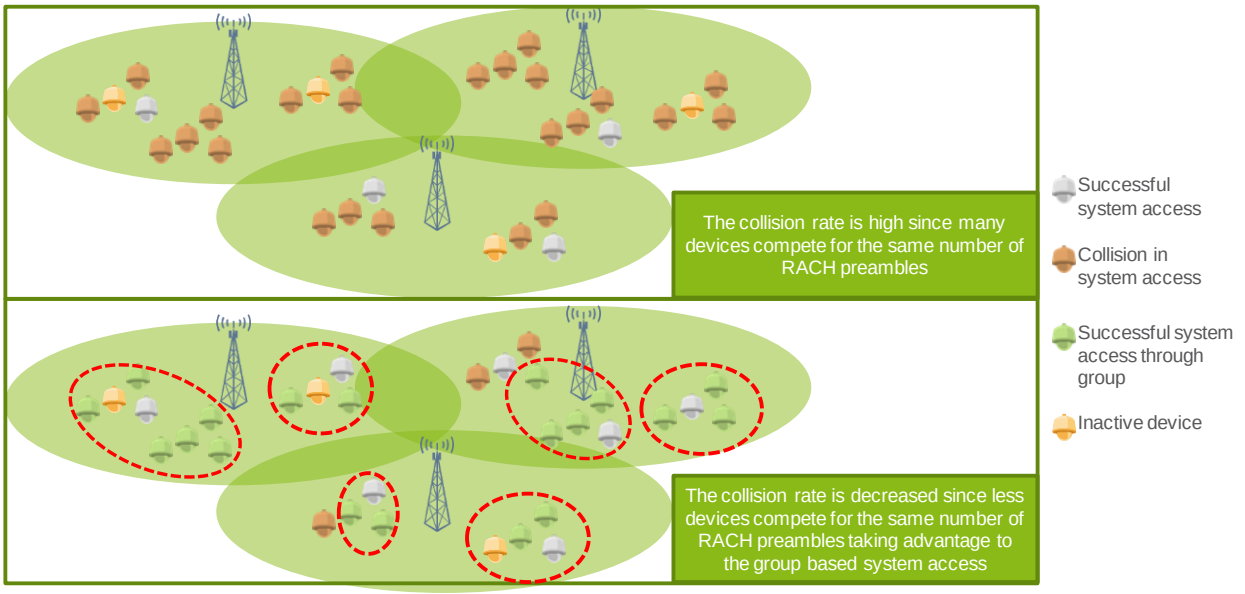


Figure 6-1: Schematic representation of the group based cluster-based RACH access

6.3 Mobility Management

Handover framework is the basic mobility management from the RAN point of view. Inter-5G handover takes place when D2D UEs move across different 5G RANs and AIVs, as shown in Figure 6-2. Intra-5G handover happens when D2D UEs move within the same 5G RAN. 5G RAN can switch D2D communications to small cells or other AIVs within the coverage of 5G RAN.

A general handover procedure includes (1) signal quality measurement, (2) coordination between the source and target BSs, (3) resource allocation of the target BS, and (4) packet switch from the source BS to the target BS. Thus, signal quality measurement is the first step in terms of the RAN-level mobility management. The D2D UEs measure link quality of ongoing communication link and possibly of nearby AIVs as indicated by RAN. To satisfy 5G RAN lean design and to reduce the “always on” system access, the measurement targets info can be carried and announced by “on-demand” SIB or dedicated signaling. “On-demand” SIB announces the common candidates of measurement targets. The dedicated signaling informs the specific configuration and candidates of measurement targets. The D2D UEs measure D2D signal and Uu signal when needed and being informed using “on-demand” SIB and/or dedicated signaling to save network power consumption. Thus, the timing as well as target of measurement can be decided by 5G RAN using the “on-demand” approach. Optionally, UE may proactively send such reports.

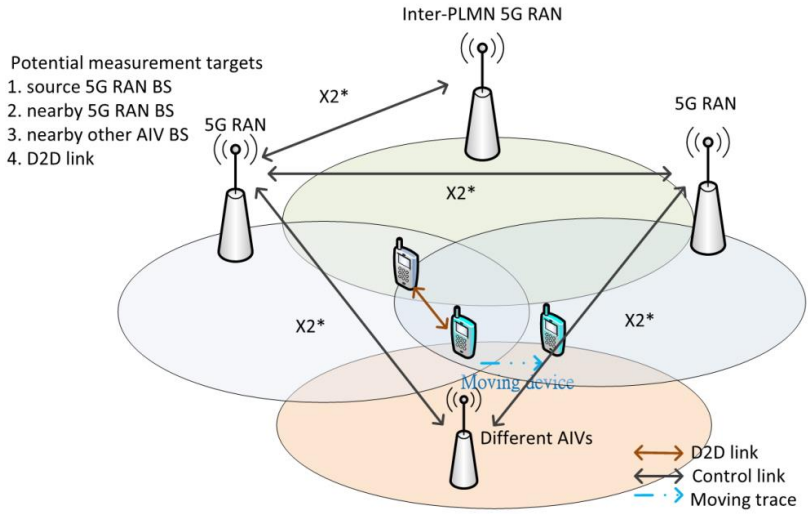


Figure 6-2: Inter-5G D2D communications

5G RAN may also communicate with RAN of other AIVs, other intra-PLMN 5G RAN and other inter-PLMN 5G RAN. Enhancements to X2* interfaces are required to support this in order to achieve inter-RAT D2D compatibility and service continuity, in terms of the second step for RAN-level mobility management, as mentioned above in the second paragraph. As such, 5G RAN requires multi-AIV signal measurement and D2D link measurement. Furthermore, in order to ensure service continuity across different RANs/AIVs, 5G RAN design can employ unified and global D2D ID for each corresponding D2D UE that enables the coordination and UE identification among different RANs/AIVs. In such cases where D2D UE has unique global ID, 5G RAN is able to uniquely identify all D2D UEs within a D2D group.

With proper multi-AIV signal measurement and D2D signal measurement, handover decisions can be smartly made to enhance D2D service continuity. D2D signal measurement is done using layer reference signal e.g. similar to those used in LTE for Uu UL/DL measurement, however, in this case reference signal is exchanged over PC5* interface. After the ongoing D2D UEs send measurement report to the source 5G RAN, the source 5G RAN is able to perform handover decision and determine the type of D2D handover that the D2D UEs should follow. To be more specific, the measurement report includes information such as radio link signal quality of the candidate measurement targets (target 5G RAN), radio link signal quality of the source 5G RAN, and radio link signal quality of the ongoing D2D link. One of the reasons that 5G RAN (BS) acts as the entity to make handover decision is that 5G RAN has to find suitable target 5G RAN/different AIVs via coordination through different X2 interfaces (e.g., X2-inter-5G, X2-5G and X2-diff AIVs). Another reason is that the signaling overhead increases if the D2D UE performs handover decision. This is because each D2D UE will need to communicate with other candidate BSs and also coordinate with other communicating D2D UEs, in order to achieve simultaneous/sequential D2D handover.

After access point to which UE is currently connected to i.e. the source 5G RAN, decides target 5G RAN that the D2D UEs should handover to, the “make before break” rationale should be adopted. This rationale ensures that the D2D context information is transferred from the source 5G RAN to the target 5G RAN before actual handover execution. Context information here refers to D2D specific information such as communication types, radio resources, traffic type, D2D group, etc. thus enabling the target 5G RAN to prepare customized D2D resources for the D2D unicast or group-cast communication based on the received D2D context. The reason of doing so is that D2D groups have distinct service requirements due to variety of applications and scenarios, and a unified service cannot fit the requirements of various types of D2D groups and communications. For example, an ongoing D2D communication group might be in a star topology with delay-sensitive traffic in a unicast communication manner. The D2D context can be obtained via series of inputs from UEs. All context information may be kept in a centralized Mobility Management function or entity and may be located in the RAN or core network.

Finally, after the handover decision completes and handover command is sent to D2D UEs, they execute handover and might use original, temporary or newly assigned D2D resources to continue the ongoing D2D communications. The handling of D2D resources across different AIVs is another issue. One approach may be that all D2D UEs use the original D2D resources for D2D communications until all D2D UEs hand over to the new target RAN/AIVs. This approach makes D2D communications continuous when D2D UEs perform the handover procedure. Another way is to design temporary D2D resources used for D2D UEs performing the handover procedure. The temporary D2D resources, besides the D2D resources provided by the original RAN or by the target RAN/AIVs, can be only used by D2D UEs performing handover procedure for data transmission.

Another aspect of mobility management and signaling optimization is related to group based design, for example, by exploiting various existing grouping schemes and secondary interfaces/connections for handling certain control operations. An approach is 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) TAL, and not predefined ones. D2D communication or secondary interfaces (e.g., 802.11p, 802.15, etc.) may be used to form groups and to exchange information within the group. According to the proposed state handling this would be configured for UEs in RRC Connected (Inactive) and/or RRC Idle. The group representative could be either in RRC connected-inactive or idle state and perform TAU, measurement reports, resource allocation and other information e.g. related to handover procedures, on behalf of the group. In RRC connected state the network automatically knows his location. In both cases the location of all the group members is being updated. Compared to the legacy systems – without group communications – where all devices should perform TAU when they are crossing the TAL borders, only the group representative will perform TAU on behalf of the group. This enables the reduction of the size of the paging areas’ size in an efficient manner. This approach enables on the one hand performing location update (when the group head is not connected) more often for

the overall group, even on a per cell basis and on the other hand targeted paging in the cells where the UE is located thus significantly reducing the latency for locating a user. In the group based approach once a device/UE joins a group then the network is 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 needs to be considered. Also, the tradeoff among the gains of the TAU suppression and the group formation and maintenance needs to be further investigated. However, the proposed scheme with the reduced user oriented TALs enables significant reduction of the paging area without potential paging misses/failures.

6.4 D2D for mMTC

As one of the new emerging services, mMTC is considered as an important service in 5G networks. Here exploitation of D2D communication to improve availability and reduce transmit power of mMTC devices is introduced. Signaling diagrams related with D2D grouping and communication procedures are provided in Annex A.1 to illustrate the proposed solution. Moreover, in order to optimize the system performance, the proposed solution also enables exploitation of context information with network assistance for D2D.

Since LTE network was not designed for mMTC service, some solutions are proposed in the literature in order to compensate for the degradation of receiver performance. For instance, in order to maintain the receiving SINR, exploiting massive TTI bundling [3GPP15-151948] [SM09] or narrow band transmission NB-IoT [3GPP15-45820] are proposed in 3GPP. The subcarrier bandwidth in NB-IoT technique is chosen to meet the system coupling loss objective given the maximum MS transmit power. However, as the maximum data rate per subcarrier is reduced in accordance with the subcarrier bandwidth, per packet transmission duration will be increased and thus the proposed solutions experience huge resource usage at system level and battery drain at device level, which can significantly reduce battery life of the device.

.

In this work, D2D communication is exploited to offer a more power and resource efficient mechanism to support wide deployment scenarios for mMTC service, including deep coverage. Deep coverage refers to the case where mMTC devices are deployed deep inside a building and experience a very challenging radio propagation scenario with a large penetration loss. It is assumed further that all machine sensors are static and they have the ability to serve as relays. In the proposed solution, certain mMTC UEs are selected to act as relay UEs by a center entity, i.e. base station, with assistance of its gathered context information. Afterwards mMTC UEs located in cell boarder or deep-in-door can set up a sidelink with their assigned relay UEs. Thus, power consumption and coverage for the remote mMTC UEs can be improved.

We assume that all sensors can receive the control information from their served BS in downlink and the main challenge to handle is the uplink transmission. In other words, remote mMTC UEs cannot reach the BS in uplink, no matter in CP or UP. Below are the main aspects of this scheme:

- Context information will be collected and exploited by BS to achieve a smart resource management algorithm which improves system performance in terms of coverage and power consumption. For instance, if certain sensors experience good channel propagation for their cellular links and their remaining battery power is sufficient for providing relay service, BS can command them to act as sensor relays.
- The sidelink group decision should be dedicated by BS to corresponding sensors.
- Deep-in-door or cell edge mMTC UEs can transmit their user package to BS through their dedicated relay sensors.

In Annex A.1, the following four procedures are stated to enable using context aware sidelink communication for mMTC services, as:

- Initial remote UE grouping - once a remote UE is attached to a network, initialization is required where BS will assign a relay UE and a sidelink group will be set up.
- Update of sidelink pairs - in case certain conditions are not fulfilled for sidelink any more, corresponding sidelink groups will be released and BS will perform its context-aware sidelink grouping algorithm again and reset the sidelink groups.
- User data transmission with sidelink communication - in case certain remote UEs are directly paged by network by listening to PCH or if they have data packets to transmit in uplink, the formed sidelink group as stated previously will be used to relay the data from remote UEs.
- Sidelink monitor and release - once a sidelink group is established, both sidelink ends and BS will monitor the conditions for sidelink and decide to release the sidelink when it is necessary, i.e. due to a low battery level of relay UE or a high propagation loss of sidelink.

6.5 Cooperative D2D

Cooperative D2D communications where D2D pairs are utilized as relays to facilitate the transmission between a cellular user (CU) and its base-station (BS) is a way to improve spectrum efficiency. Thus, 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). Besides, such D2D devices facilitate cellular user transmission by acting as Relay device, as shown in Figure 6-3. Cooperative communication scheme as proposed here 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. Such cooperative modes of communication as discussed here is unknown to LTE-A however, considering its many benefits and much ongoing work in this direction it is likely to be an important aspect of D2D communication in 5G.

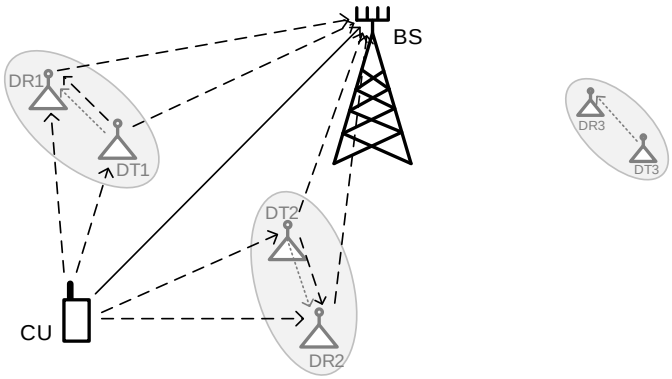


Figure 6-3: Illustration of multi-relay cooperative D2D communications

To enable cooperative D2D communications, here we discuss approaches for cooperative mode selection, relay selection, cooperative transmission, and resource allocation. Several issues related to cooperative D2D are discussed in Annex A.3. In the following, we focus our discussions on the uplink whereas the downlink can be done similarly. However, in the downlink, one can further discuss the power allocation among different channels and data streams.

Mode selection (addition of cooperation modes and mode selection policies to incorporate cooperation):

In the context of cooperative D2D communications, mode selection refers to the BS decision on whether a D2D link should adopt the underlay, overlay, or hybrid cooperative relay mode.

Specifically, in the cooperative relay mode, we propose three types of cooperation: overlay, underlay, and hybrid cooperation, as described below, and their frame structures are illustrated in Figure 6-4.

- In underlay cooperation, the transmission is divided into two phases: a direct transmission phase and a (superposition) cooperative transmission phase. In the direct transmission phase, CU transmits its data directly to the BS, DTs, and DRs; then, in the cooperative transmission phase, DTs will transmit a signal consisting of the superposition of both CU's data and the data intended for their respective DRs. The DRs will decode using Success Interference Cancellation (SIC) techniques, if the CU's message is decodable, and will simply treat it as interference, otherwise.
- In overlay cooperation, the transmission is divided into three phases: a direct transmission phase, a clean-relaying phase, and a D2D transmission phase. The clean-relaying phase is dedicated to the relaying of CU's messages (by both DTs and DRs) whereas the D2D transmission phase is dedicated to the transmission between DTs and DRs.
- The hybrid cooperation mode is the combination of the two where superposition relaying is used in the second phase of overlay mode instead of clean relaying.

  The above cooperation modes can be selected in addition to the usual overlay or underlay mode used by D2D UEs without any cooperation.

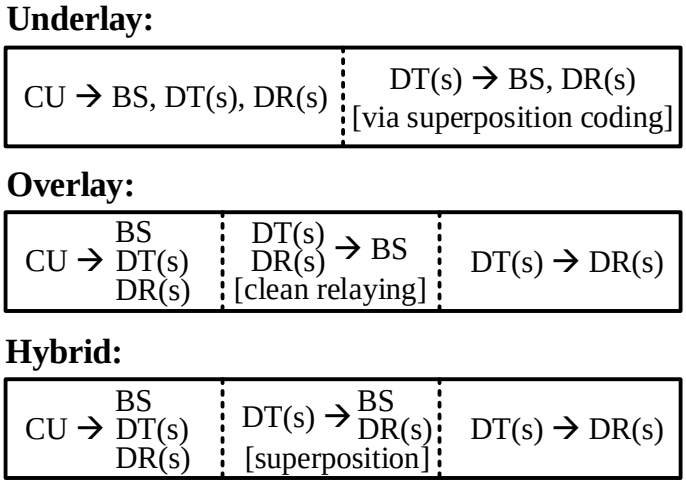


Figure 6-4: Illustration of overlay, underlay, and hybrid cooperation modes.

By taking into consideration the cooperation mode, the complexity of centralized mode selection increases considerably since one must take into account the strong dependence between the relay selection and cooperative mode selection, as well as the co-channel interference between different D2D pairs and cellular UE that are admitted in the same time-frequency channel. To simplify the mode selection process and to increase its compatibility with systems that do not employ D2D cooperation, we propose a multi-stage mode selection and channel assignment policy as follows.

- In stage 1, mode selection and channel assignment is first performed without consideration of the possible cooperation from D2D pairs. That is, CUs and D2D pairs are assigned the same time-frequency channels only if the co-channel interference is sufficiently small between them, i.e. a CU is far away from a D2D pair.
- In stage 2, D2D pairs that were not assigned in stage 1 will seek to relay for one of its candidate CUs in order to further increase its spectrum opportunity. A D2D pair will cooperate only if its throughput or reliability can be improved by relaying for the cellular user. By allowing multiple D2D pairs to help one CU, the amount of resources required for each D2D pair to cooperate may be reduced, leaving more resources for its own transmission.

To enable cooperation, the BS not only needs to know the channel between itself and the users, but also need to know the inter user channels, namely, the channels between CUs and D2D users as well as the channel between D2D pairs. Nonetheless, by having each D2D pair determine a local set of candidate CUs, the number of channel parameters from CUs to D2D pair that needs to be conveyed to BS is reduced.

Selection of candidate CUs for cooperative D2D transmission

Different from many works on cooperative relaying that discuss the selection of relays to enhance the transmission of the source node, we propose a D2D-initiated cooperation where D2D pairs instead look for CUs that they are willing to cooperate with. In particular, a D2D pair will include a CU into its candidate set if the probability that the D2D rate is achievable while relaying for the

CU at its (discounted) target rate is beyond a certain threshold. The outage probability threshold can be chosen to limit the size of the candidate CU set. A discount on the target rate of the CU is adopted so that a CU can be included in a D2D pair's candidate user set even if the D2D pair is not able to help achieve the target rate by itself, but can do so with the help of other D2D pairs.

RAN could assist in selecting a set of suitable CUs. In this case, the relayed connection is triggered by the eNB and the selection of the most suitable CUs is performed by the eNB which in turn commands the Relay UE(s) to trigger the CU-Relay connection setup procedure. Annex A.3 provides further details of CU selection procedure.

Interference-aware cooperative transmission schemes

With multiple relays, several distributed MIMO techniques, such as distributed space-time coding, opportunistic relay selection, and distributed beamforming, can be adopted depending on the level of synchronization that can be achieved among the D2D pairs and the channel state information available at the destination. The cooperative transmission schemes adopted affects the rates that can be achieved by both the cellular and D2D links, and also the decision in the mode selection process. When the BS is equipped with multiple antennas, receive and transmit beamforming in the uplink and downlink, respectively, must also be carefully designed to fully exploit the cooperative advantage.

In particular, opportunistic relay selection is a promising technique in practice since it does not require strict synchronization among relay nodes if only one or a few are selected. More specifically, the idea is to choose one (or multiple) relays with the best instantaneous channel (or channels) among those that are cooperating to forward the information. This technique is known to achieve the maximum diversity order in interference-free environments. Distributed space-time coding or distributed beamforming can be used in conjunction with the relay selection if more than one relay is selected. In D2D environments, relay selection should further take into consideration the following points:

- Interference is a performance dominating factor in D2D networks and should be taken into consideration in the relay selection policy. That is, the relay performance of a D2D pair may be affected by the interference it receives from other active D2D pairs and cellular UEs in the same channel and the selection of a D2D pair as relay may cause additional interference to other D2D pairs that are assigned to the same channel.
- D2D pairs also have their own data to transmit and, thus, the relay selection should take into consideration the throughput requirement of each D2D link as well as the total throughput improvement that can be provided.
- The selection of each D2D pair yields 2 potential relays (i.e., DT and DR) for the source (i.e., CU in the uplink or BS in the downlink).

Furthermore, when underlay or hybrid cooperation modes are adopted, superposition coding is used. That is, the relay and the D2D messages are superimposed in the cooperative transmission phase and successive interference cancellation is required at DRs. In this case, relaying by multiple DTs can be helpful not only in terms of providing spatial diversity to the cellular link, but can also provide diversity to the interference forwarding (i.e., forwarding of the cellular signal to

the DRs) to improve DR's capability to decode and cancel the strong interference signal. This should be taken into consideration in the cooperative transmission design and the corresponding power allocation.

6.6 D2D Network Offloading

Content caching near the consumer is a way for enhancing QoS, E2E latency and traffic offloading, thus enabling to meet key performance indicators or KPIs e.g. related to use case 1 [MET16-D11]. Caching at the base stations falls within this concept. There are ongoing works that proposes to use helper stations (femto caching) that store most popular video files, and transmit them, upon request, via short-range wireless links to the user terminals. Combining D2D with in-network caching exploits advantages of D2D communications for a fast content retrieval. There is a huge D2D potential from a general network offload point of view, considering a D2D enabled mobile network where users retrieve, in priority, contents from neighbors' caches using single-hop and two-hop D2D communication. Otherwise, if the device does not find the desired content among nearby devices, it will be downloaded directly from the Internet through the base station. Here focus is on the offload potential from cellular network to D2D interface.

Implementation of such a mechanism needs special considerations. In a classical IP-based mobile network operation, a device requests the content from the server. Whereas in the proposed approach, the base station intercepts this request and verifies if this content is present in one or more devices in its coverage area. The base station then communicates to the requesting device the list of these caches and the latter verifies if one or more of these caches is within its list of D2D neighbors. If yes, it initiates a direct D2D communication for retrieving the content. Otherwise it requests from the server.

7 Summary

The METIS-II project aims at developing a comprehensive and detailed 5G RAN design in order to support the standardization. This deliverable describes the mid-point view of the asynchronous control functions and overall control plane design. This includes initial conclusions on the design of the Radio Resource Control (RRC) protocol for the 5G RAN, as well as for the asynchronous CP functions initial access, RRC state handling, mobility and native support of D2D. The asynchronous CP functions and the CP architecture must be able to fulfill a wide variety of requirements such as futureproof design, high energy efficiency, beamforming mobility and higher connection reliability. This deliverable mainly focuses on how the asynchronous CP functions can handle these requirements. Regarding the overall CP design, an important aspect highlighted is the ability to have tight inter-working with LTE-A evolution from the beginning in order to be able to have a smooth migration of new 5G networks, together with the standalone operation of the 5G networks. This puts several requirements on the CP architecture and the RRC design which is addressed in this deliverable. Additionally, several technical aspects related to initial access, mobility and D2D connectivity of 5G are investigated. All the afore considerations will be further analyzed throughout the lifetime of the project and reported in the following deliverable.

8 References

[3GPP11-22801] 3GPP TR 22.801, “Study on non-MTC mobile data applications impacts (Release 12)”, Rel. 12, V.12.0.0, Dec. 2011.

[3GPP11-36321] 3GPP TS 36.321, “Evolved universal terrestrial radio access (E-UTRA); medium access control (MAC); protocol specification”, Rel. 10, V.10.2.0, Jun. 2011.

[3GPP16-36331] 3GPP TS 36.331, “Evolved universal terrestrial radio access (E-UTRA); radio resource control (RRC); protocol specification”, Rel. 13, April 2016.

[3GPP11-36822] 3GPP TR 36.822,“LTE RAN enhancements for diverse data applications”, V0.2.0, Nov. 2011.

[3GPP13-36912] 3GPP TR36.912 version 13.0.0 Release 13

[3GPP14-38913] 3GPP TR 38.913 version 0.3.0 Release 14

[3GPP15-23060] 3GPP TS 23.060 version 13.6.0 Release 13

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

[3GPP15-150698] 3GPP TSG SA WG2 Meeting #107, S2-150698.

[3GPP15-150946] 3GPP TSG-RAN3 Meeting #88, R3-150946.

[3GPP15-151948] 3GPP meeting, RP-151948, “New WI Proposal: D2D based MTC”, December 2015

[3GPP15-22720] 3GPP TR 22.720, “Architecture Enhancements for Cellular Internet of Things”, Release 13, 2015

[3GPP15-22891] 3GPP TR 22.891, “Feasibility Study on New Services and Markets Technology Enablers”, Release 14, 2015

[3GPP15-23303] 3GPP TS 23.303, Proximity-based services (ProSe); Stage 2, 2015.

[3GPP16-23799] 3GPP TR 23.799, ” Study on Architecture for Next Generation System”, http://www.3gpp.org/ftp//Specs/archive/23_series/23.799/23799-050.zip

[3GPP15-25303] 3GPP TS 25.303, “Interlayer procedures in Connected Mode (Release 13)”, December 2015

[3GPP15-36300] 3GPP TS 36.300, “E-UTRAN; Overall Description; Stage 2; (Release 12)”, Sept. 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

[3GPP15-RP-151621] 3GPP, RP-151621, “New Work Item: NarrowBand IOT (NB-IOT)”, September 2015

[3GPP16- 36311] 3GPP TS 36.311, Evolved Universal Terrestrial Radio Access (E-UTRA); Radio Resource Control (RRC); Protocol specification

[3GPP16-36885] 3GPP TR 36.885: "Study on LTE-based V2X Services", March 2016

[3GPP16 R1163961] 3GPP R1-163961, TSG RAN WG1 Meeting #85, “Final Report of 3GPP TSG RAN WG1 #84bis v1.0.0”, April 2016

[3GPP16-R2-163441] 3GPP R2-163441, “Discussion of RRC States in NR”, 3GPP TSG-RAN WG2 Meeting #94, Nanjing, China, May 2016

[3GPP16-RP152293] RP-152293, “New WI proposal: Support for V2V services based on LTE sidelink”, LG Electronics, Huawei, HiSilicon, CATT, CATR, Dec. 2015.

[3GPP16-RP160540] 3GPP Working Item Description: “Signalling reduction to enable light connection for LTE” RAN2, REL-14, March 2016

[3GPP16-RP160649] RP-160649, “Revised WID: Support for V2V services based on LTE sidelink”, LG Electronics, Huawei, HiSilicon, CATT, CATR, March 2016

[3GPP16-S2161323] 3GPP S2-161323, “Solution: Mobility Framework”, 3GPP SA2 WG, Sophia Antipolis, France, Feb. 2016

[3GPP16-TR23720] 3GPP TR 23.720, “Study on architecture enhancements for Cellular Internet of Things, (Release 13)”, March 2016

[ALC13] Alcatel Lucent Application Note “The Impact of Small Cells on MME Signaling, Methods to Reduce and Optimize MME Core Signalling Caused by Small Cells”, October 2013.

[BHM] N. Barati, S. Hosseini, M. Mezzavilla, S. Rangan, T. Korakis, S. Panwar, and M. Zorzi, “Initial Access for Millimeter Wave Cellular Systems”, submitted to IEEE Transactions on Wireless Communications.

[BHM+15] N. Barati, S. Hosseini, M. Mezzavilla, Parisa Amiri-Eliasi, S. Rangan, T. Korakis, S. Panwar, and M. Zorzi, “Directional Initial Access for Millimeter Wave Cellular Systems”, in Proc. Asilomar Conf. on Signals, Systems and Computers, 2015.

[BHR+15] N. Barati, S. Hosseini, S. Rangan, P. Liu, T. Korakis, S. Panwar, and T. Rappaport, “Directional cell discovery in millimeter wave cellular networks,” IEEE Transactions on Wireless Communications, vol. 14, no. 12, pp. 6664–6678, 2015.

[BPR+15] S. Barbera, K. Pedersen, C. Rosa, P. Michaelsen, F. Frederiksen, E. Shah, A. Baumgartner, Synchronized RACH-less Handover Solution for LTE Heterogeneous Networks, International Symposium on Wireless Communication Systems. IEEE Press, 2015

[Chu11] Yun Won Chung, “Adaptive design of tracking area list in LTE”, Wireless and Optical Communications Networks (WOCN), 2011 Eighth International Conference on , vol., no., pp.1,5, 24-26 May 2011.

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

[CKS+15] Konstantinos Chatzikokolakis, Alexandros Kaloxylos, Panagiotis Spapis, et al., On the Way to Massive Access in 5G: Challenges and Solutions for Massive Machine Communications, CrownCom 2015.

[CSW15] Ho-Yuan Chen, Mei-Ju Shih, Hung-Yu Wei, “Handover Mechanism for Device-to-Device Communication” IEEE CSCN'15.

[CWH+13] Fu-Min Chang, Hsiu-Lang Wang, Szu-Ying Hu, Shang-Juh 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, 2013, pp 270-280

[ERI14] Ericsson Mobility Report, On the Pulse of the Networked Society, November 2014

[FAN5G] FANTASTIC 5G project (“flexible air interface for scalable service delivery within wireless communication networks of the 5th Generation”), see http://fantastic5g.eu//http://fantastic5g.eu//

[GMR+16] M. Giordani, M. Mezzavilla, S. Rangan, M. Zorzi, "Multi-Connectivity in 5G mmWave Cellular Networks", to appear in IEEE MED-HOC-NET 2016.

[GSMA+16] GSMA; https://www.gsmaintelligence.com/, 2016-06

[JSR+06] M. Johnsson, J. Sachs, T. Rinta-aho, T.Jokikyyny, “Ambient networks – a framework for multi-access control in heterogeneous networks”, in proc. IEEE Vehicular Technology Conference (VTC), Montreal, Canada, Sep 2006.

[KKP+13] Younghyun Kim, Haneul Ko, Sangheon Pack, Wonjun Lee, and Xuemin (Sherman) 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, H.D. Schotten, Exploiting Diurnal User Mobility for Predicting Cell Transitions, GLOBECOM WS, Atlanta, USA, 2013

[KKS15] N.P. Kuruvatti, A.Klein, H.D. Schotten, Prediction of Dynamic Crowd Formation in Cellular Networks for Activating Small Cells, VTC-Spring WS, Glasgow, Scotland, 2015

[LH05] S.Liou and Y.Huang, Trajectory Predictions in Mobile Networks, International Journal of Information Technology Vol.11, No.11, 2005

[Maz75] J. E. Mazo, “Faster-than-Nyquist signaling”, Bell System Technical Journal, vol. 54, no. 8, pp. 1451-1462, October 1975.

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

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

[METIS] ICT-317669 METIS project, see http://metis2020.com

[MET13-D11] ICT-317669 METIS, Deliverable 1.1 Version 1 “Scenarios, requirements and KPIs for 5G mobile and wireless system”, April 2013.

[MET13-D61] METIS, D6.1 - Simulation guidelines, October, 2013, available online: https://www.metis2020.com.

[MET16-D11] METIS II, Deliverable 1.1 Version 1 “Refined scenarios and requirements, consolidated use cases, and qualitative techno-economic feasibility assessment”, January 2016.

[MET16-D22] ICT-671680 METIS-II, Deliverable D2.2, “Draft Overall 5G RAN Design”, June 2016

[MET-II16-D51] METIS-II, Deliverable D5.1, “Draft Synchronous Control Functions and Resource Abstraction Considerations”, May 2016

[MKS+10] A. Merentitis, A. Kaloxylos, M. Stamatelatos, N. Alonistioti, "Optimal periodic radio sensing and low energy reasoning for cognitive devices," 2010 15th IEEE Mediterranean Electrotechnical Conference (MELECON 2010), April 2010

[mmMAGIC] mmMAGIC project, see https://5g-ppp.eu/mmmagic/https://5g-ppp.eu/mmmagic/

[NGM15] NGMN Alliance, “5G White Paper”, Feb. 2015, available at http://www.ngmn.org/fileadmin/ngmn/content/images/news/ngmn_news

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

[PF13] Monica Paolini, Senza Fili , The taming of the app Measuring the financial impact of mobile traffic optimization, 2013.

[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, Y. He., "Cell selection analysis in outdoor heterogeneous networks", Proc. of IEEE ICACTE, 2010

[RVM12] D. Ramasamy, S. Venkateswaran, and U. Madhow, “Compressive tracking with 1000-element arrays: a framework for multi-gbps mm wave cellular downlinks,” in Proc. 50th Ann. Allerton Conf. on Commun., Control and Comp., Monticello, IL, Sep. 2012.

[SM+15] I. Da Silva, G. Mildh at al., “Tight integration of new 5G air interface and LTE to fulfill 5G requirements,” Vehicular Technology Conference (VTC Spring), IEEE 81st, pp. 1-5, Glasgow, May 2015.

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

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

[MSC13] P. Makris, N. Skoutas, C. Skianis, A Survey on Context-Aware Mobile and Wireless Networking: On Networking and Computing Environments’ Integration, IEEE Communication Surveys and Tutorials, 2013

[SURDIM] Imola K. Fodor. A survey of dimension reduction techniques. 2002.

[SWY+10] Hongbo Si, Yue Wang, Jian Yuan, Xiuming Shan, Mobility Prediction in Cellular Network Using Hidden Markov Model, Consumer Communications and Networking Conference (CCNC), 2010.

[TAZ+13] A. Tzanakaki, M. P. Anastasopoulos, G. S. Zervas, B. R. Rofoee, R. Nejabati and D. Simeonidou, “Virtualization of heterogeneous wireless-optical network and IT infrastructures in support of cloud and mobile cloud services”. IEEE Communications Magazine, vol. 51, no. 8, pp. 155-161, August 2013.

[TCG+14] A. Talukdar, M. Cudak, A. Ghosh, Handoff Rates for Millimeterwave 5G Systems, Vehicular Technology Conference (VTC), 2014

[TP10] T. Thumthawatworn, A. Pervez, "Multi-level rule-based handover framework for heterogeneous wireless networks," 6th Conference on Wireless Advanced (WiAD), June 2010

[XCG06] Y. Xiao, H. Chen, and M. Guizani, “Performance Evaluation of Pipeline Paging under Paging Delay Constraint for Wireless Systems,” IEEE Trans. Mobile Computing, vol. 5, no. 1, pp. 64-76, Jan. 2006.

[XPM+14] Dionysis Xenakis, Nikos Passas, Lazaros Merakos, Christos Verikoukis, “Mobility Management for Femtocells in LTE-Advanced: Key Aspects and Survey of Handover Decision Algorithms”, Communications Surveys & Tutorials, IEEE (Volume:16 , Issue: 1), 1st Quarter 2014

[ZN13] K. Zhou, N. Nikaein, “Packet aggregation for machine type communications in LTE with random access channel,” Wireless Communications and Networking Conference (WCNC), 2013 IEEE, pp.262-267, April 2013.

[ZVM12] H. Zhang, S. Venkateswaran, and U. Madhow, “Analog multitone with interference suppression: Relieving the ADC bottleneck for wideband 60 GHz systems,” in Proc. IEEE Globecom, 2012.

A Annex

A.1 Solution to exploit sidelink for mMTC service

In this part, four procedures to enable sidelink communication for mMTC service as stated in Subchapter 6.4 are given.

A.1.1 Initial remote UE grouping

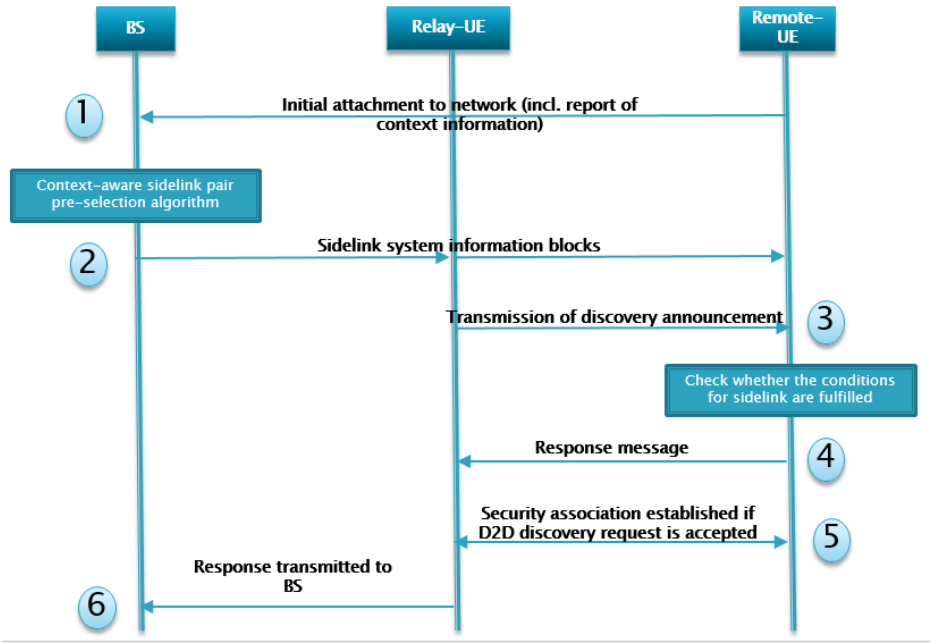


Figure A1-1: Initialization procedure of a remote UE

Once a remote UE is attached to a network, initialization is required. Figure A1-1 shows the procedure for initialization of remote UE as:

1. The UE's location and other information are reported to its serving BS when UE is initially attached to NW. BS will analyse the location, battery level, traffic type and RSRP of its connected mMTC UEs and perform the context aware sidelink pair preselection algorithm.
2. After obtaining the decision of its context aware sidelink algorithm, BS will inform UEs with sidelink system information blocks, i.e. the IDs of pre-selected remote-UEs and IDs of pre-selected relay-UEs, the resource used for sidelink discovery and communication, conditions that sidelink pairs should fulfill.
3. Relay UE should send discovery announcement message to target remote UE by containing the ID of the target UE.

4. Based on the calculated S-RSRP, target UEs could determine whether the request is accepted or not (with the help of assistance information acquired from BS). And the decision should 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, i.e. those specified in 3GPP TS 33.303, or physical layer security algorithm.

6. If the request is not accepted, a discovery reject message should be sent to the relay UE and the BS. After receiving the reject message, relay UE will avoid further request of direct connection with this remote UE, and BS will re-select the relay UE for the target remote UE and repeat the procedure from step 2 to step 5.

Configuration of relay UE in this procedure:

- The sidelink system information block provides relay UE the configuration information used for sidelink discovery.
- The sidelink system information block also informs relay UE about the configuration information used for transmitting sidelink discovery decision from relay UE to BS.

Configuration of remote UE:

- The sidelink system information block provides remote UE the configuration information used for sidelink discovery.

Notice:

- In this model the relay UE broadcasts discovery announcement messages on the resource configured by the sidelink system information block and the monitoring UEs that are interested in these messages read and process them.
- In case synchronous transmission is used for sidelink, corresponding sidelink pair requires a synchronization process.
- Once a sidelink pair is successfully attached to each other, certain context information should be stored by both the sidelink ends, even in connected-inactive state, i.e. by both relay and remote UEs.

A.1.2 Update of sidelink pairs

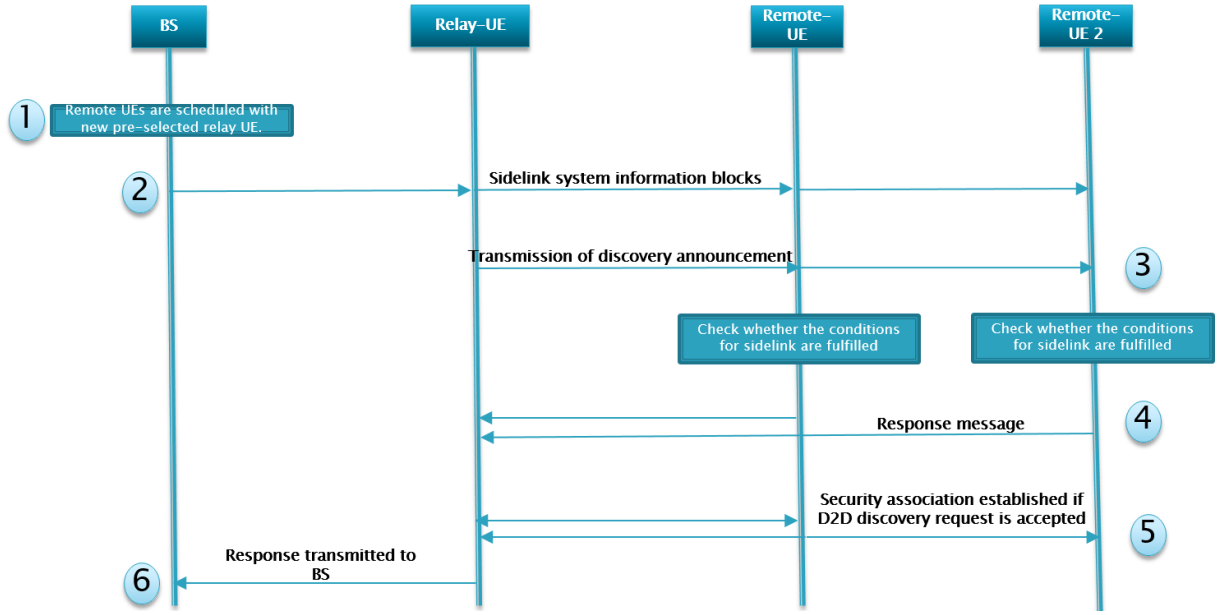


Figure A1-2: Procedure for update of sidelink pairs

In case if certain conditions are not filled for sidelink communication any more, the corresponding sidelink groups will be released and BS will perform its context-aware sidelink algorithm again and reset the sidelink groups.

Figure A1-2 shows the procedure for updating of sidelink pairs. We show here only two remote UEs for sidelink pair update as an example, in case if a group of UEs are assigned with the same relay UE.

1. BS will analyse the location, battery level, traffic type and RSRP of its served mMTC UEs and perform the context aware sidelink preselection algorithm.
2. BS will inform UEs with sidelink system information blocks for direct discovery and communication, i.e. the IDs of pre-selected remote-UEs and IDs of pre-selected relay-UEs, the resource used for sidelink discovery and communication, conditions that sidelink pairs should fulfill.
3. Relay UEs should send discovery announcement to target remote UEs by containing their IDs or the group IDs of the target UEs.
4. Based on the calculated S-RSRPs, target UEs could determine whether the request is accepted or not (with the help of assistance information acquired from BS). And the decision should 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, i.e. those specified in 3GPP TS 33.303, or physical layer security algorithm.

The result on sidelink pair update should be further transmitted from relay UE to the serving BS. If the request is not accepted, the relay UE should inform the serving BS and BS should reselect the relay UE as from step 1 to step 5 and avoid to pair the previous selected relay UE and this remote UE. Note: 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 sidelink discovery indicated by sidelink system information block.

A.1.3 User data transmission with sidelink communication

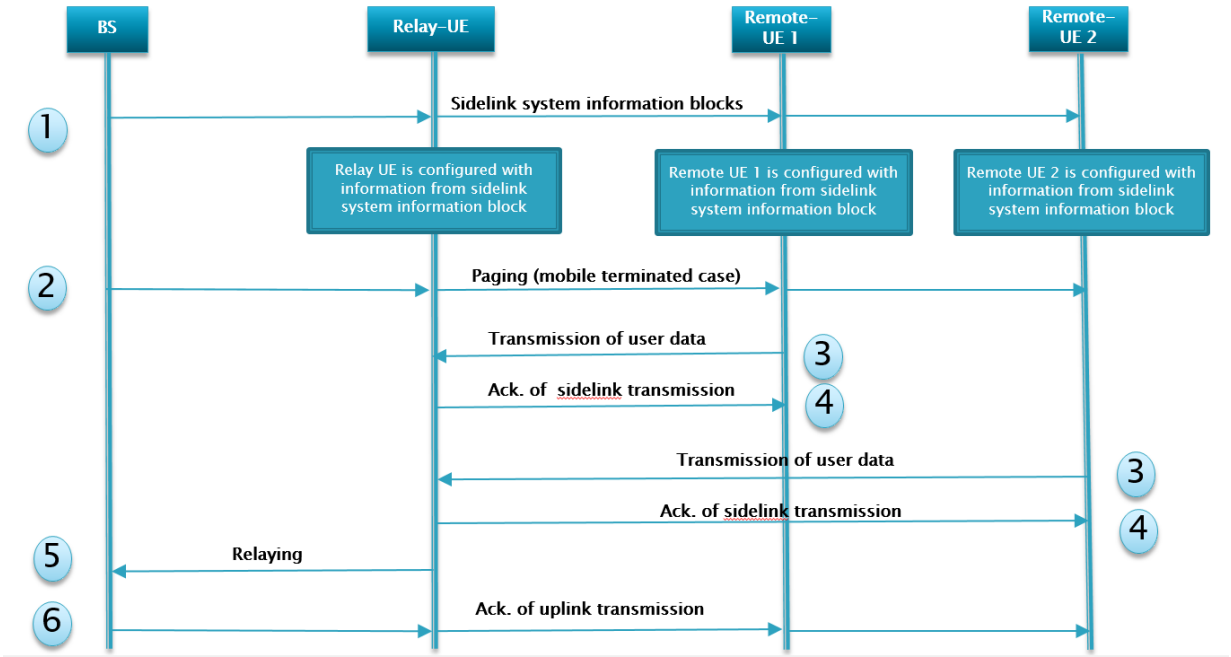


Figure A1-3: Sidelink communication procedure

In case if certain remote UEs are paged by network or if they have data packets to transmit in uplink, the formed sidelink group as stated in the previous subsection will be used to relay the data from remote UEs.

Figure A1-3 shows the sidelink communication 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.

1. Relay and remote UEs involved in sidelink communication are configured by the serving BS with sidelink system information blocks for sidelink communication.
2. In UE terminated case, one remote UE or a group of remote UEs will be paged by BS.
3. In UE originated case, a remote UE tries to transmit its data packet to its relay UE. This step also includes the sidelink random access procedure, sidelink connection setup

procedure between the relay UE and remote UE, also sidelink retransmission if certain error is experienced. Meanwhile, several information are required in this step:

- At remote UE(s): the time instance and resource that relay UE wakes up to listen to the sidelink-RACH.
- At relay UE: the resource assigned by BS for sidelink communication.

4. After successful receiving packets from remote UE(s), relay UE should respond the remote UE(s) with the acknowledgement of sidelink transmission.
5. Relay UE will forward the successful received packets to the serving BS. This process can be performed as a normal cellular uplink transmission.
6. BS will acknowledge the successful uplink transmission by sending a response message to the involved UEs.

The configuration of Relay UE includes following functions:

- Configuration information regarding the transmission of sidelink synchronization signal from relay UE (in case if relay UE acts as a synchronization reference).
- Information regarding the period when relay UE wakes up to listen to S-RACH channel.
- Information related with the resource assigned for sidelink communication.
- Information related with the relay mode, i.e. whether it is allowed for relay UE to compress the user packets from several UEs and send them once to BS.

Configuration of Remote UE:

- Information regarding when Remote UEs need to listen to the paging channel.
- Configuration information regarding when relay UE will transmit sidelink synchronization signal (in case if relay UE acts as a synchronization reference).

A.1.4 Sidelink monitor and release

Once a sidelink group is established, both sidelink ends and BS will monitor the conditions for sidelink and decide to release the sidelink when it is necessary, i.e. due to a low battery level of relay UE or a high propagation loss of sidelink. If one sidelink end decides to release the sidelink, relay UE should report this decision to BS and erase the stored context information of the paired sidelink. Figure A1-4 demonstrates the procedure of releasing a sidelink.

1. BS and sidelink ends monitor the sidelink 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 disconnect 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 accept message.

5. Sidelink ends will release and erase their stored context information of this sidelink.

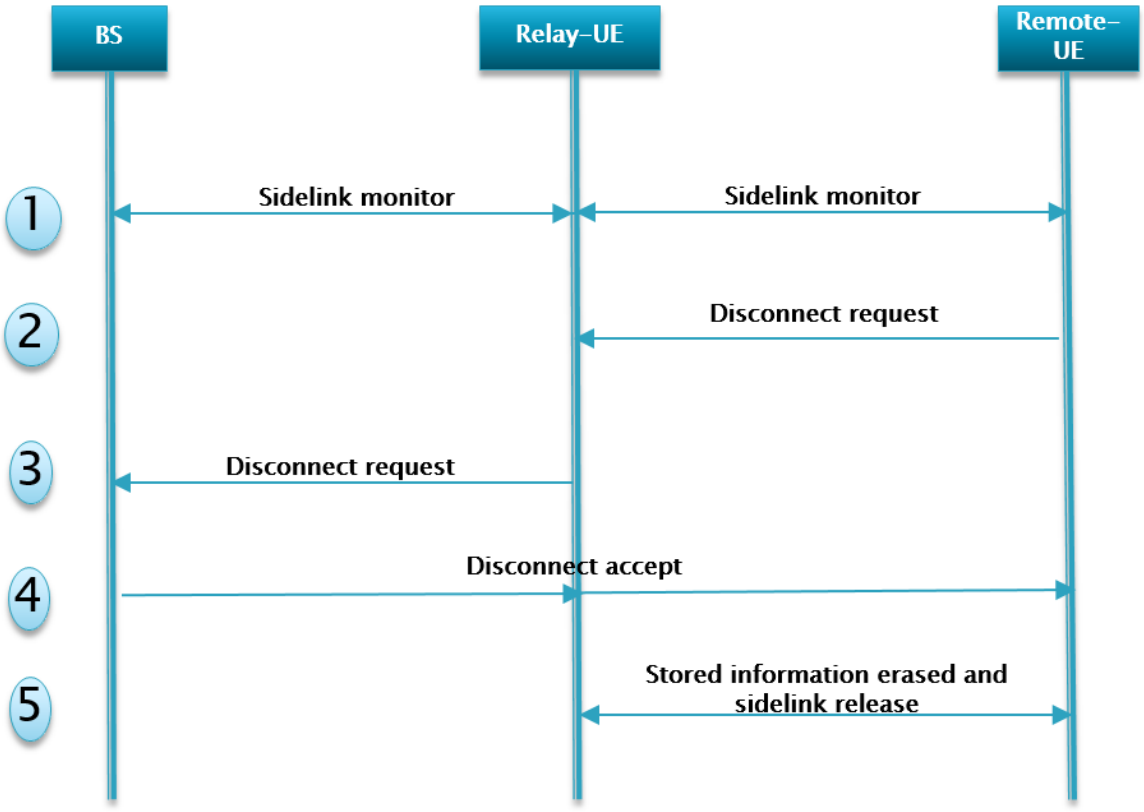


Figure A1-4: Sidelink release procedure

After the release of a sidelink, either a relay reselection procedure as stated before should be triggered in BS or the remote UE will be configured as a normal UE. The relay node previously used should be excluded when a new relay node is being selected.

A.2 Group Based Schemes

Group based schemes solution is based on below general design principles:

- Take advantage of already available groups formed for other purposes (device tracking, car to car clustering, etc.)
- Perform group location update where one device is responsible for informing the network on behalf of the group
- Having per user (or per user group) TAL, and not predefined ones.
- Take advantage of the already available interfaces for the connectivity among the nodes (e.g., 802.11p, 802.15, direct connections)

This approach will enable:

- Performing location update more often for the overall group, even on a per cell basis
- Targeted paging in the cells where the UE is located thus significantly reducing the latency for locating a user.

Figure A2-1 shows the messaging details of group based schemes. 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.

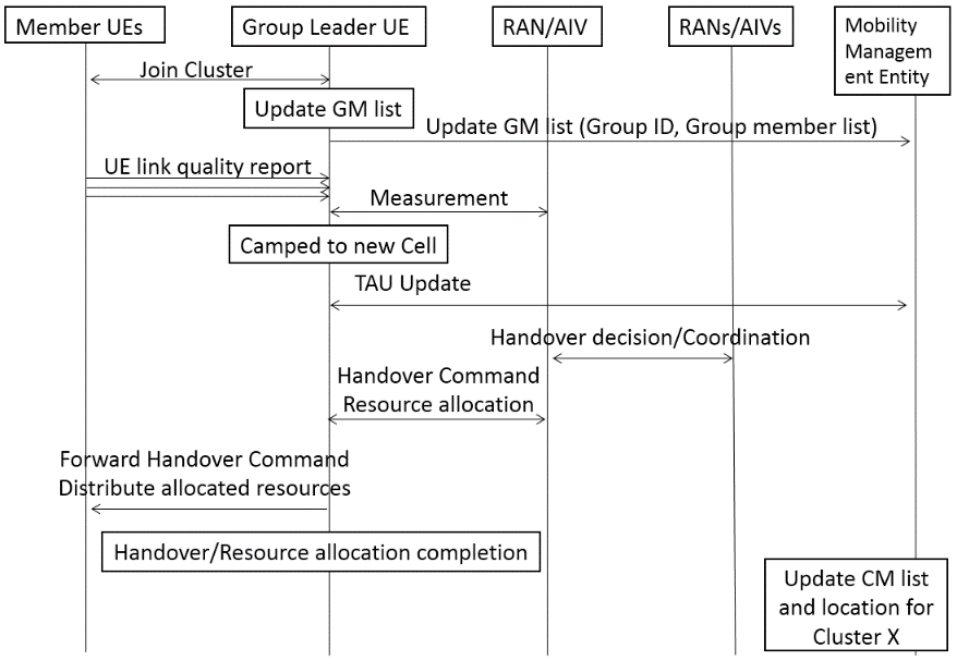


Figure A2-1: Group-head based D2D mobility management

A.3 Cooperative D2D

Advantages of cooperative D2D communications, especially with multiple nodes serving as relays shown in figure 6-3, can be summarized as follows:

- Relaying by D2D users provides D2D links with more spectrum resources to transmit their own data and, thus, increases the number of D2D connections per channel per unit area.

- The use of both DT (i.e. transmitter of the cooperating D2D pair) and DR (i.e. receiver of the cooperating D2D pair) as well as multiple D2D links helps increase the spatial diversity experienced by the cellular link, both in terms of the fading and the interference level.
- The use of multiple DTs provides spatial diversity in the interference forwarding, allowing DRs to decode the cellular interference more easily.

There are several open issues related to cooperative D2D that needs to be further investigated:

- Different cooperative modes must be determined to exploit the cooperative advantages in different scenarios.
- New mode selection mechanisms must be developed to allow for relaying by D2D users and to incorporate the different cooperative transmission modes.
- D2D pairs must identify candidate CUs in its proximity that can provide them with more spectrum opportunities through cooperative D2D communications.
- Cooperative transmission schemes that can take into consideration the transmission requirements of D2D links and can also cope with the interference in D2D environments should be developed.
- Joint mode selection and resource allocation policy should be devised to maximize the throughput or number of D2D connections per channel per unit area. That is, it is important to determine which CU is to cooperate with which D2D pair (or pairs), using which cooperation mode, and on which channel.

A.3.1 Cellular User Selection

Figure A3-1 depicts a possible message flow of CU selection procedure. A Relay UE (DR, DT) that seeks a CU for cooperation to increase its achievable rate may broadcast a Relay UE Announcement Message. Candidate CUs that are willing to cooperate send a Measurement report to RAN and a Response message to the Relay UE. The Relay UE sends a Measurement report to RAN and in turn RAN selects a set of suitable CUs and informs the Relay UEs about its selection.

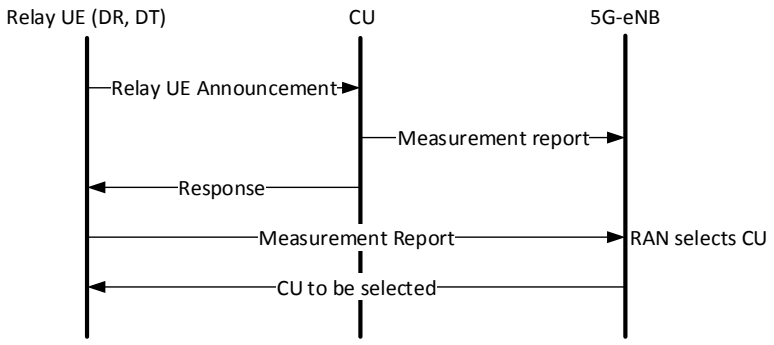
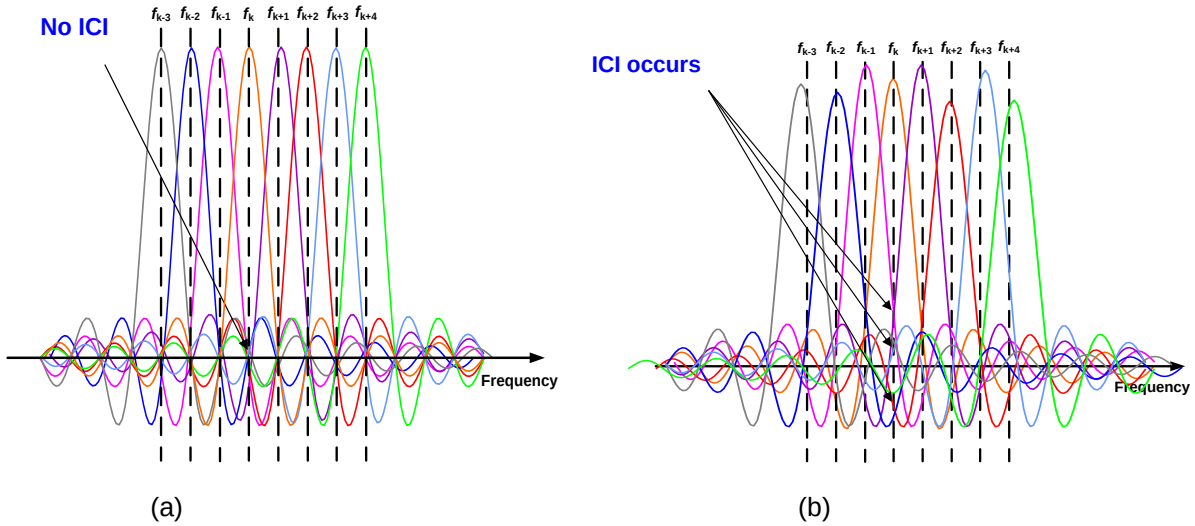


Figure A3-1: Example for CU selection procedure.

A.4 High Speed Mobility

An Illustration of OFDM received signals without Doppler effect and with Doppler effect is shown in the Figure A4-1(a) and Figure A4-1(b), respectively.



(a) (b)

Figure A4-1: An Illustration of OFDM received signals (a) without Doppler effect, (b) with Doppler effect

Conventional Doppler effect cancellation methods can be divided into the following categories:

1) ICI self-cancellation: the redundancy is used to perform the ICI self-cancellation. For example, the guard interval (i.e. cyclic prefix) inserted in the front of the OFDM signal in LTE can be used to self-cancel ICI. In another example, large redundancy is adopted to increase the ability of ICI self-cancellation. However, the redundancy will reduce the spectrum efficiency and the performance is restricted by the large normalized Doppler frequency.

2) ICI reconstruction and cancellation: parallel interference cancellation (PIC) and successive interference cancellation (SIC) in the frequency domain equalization can be used to mitigate ICI. The PIC detects all the data and then reconstructs/cancels all the interference in parallel. The SIC adopts iterative ICI cancellation, it detects the data with strongest SINR, reconstruct/cancel the interference caused by the data, and then detects other data with strongest SINR, and so on. Compare to PIC, the performance of SIC is better, however, heavy computational complexity is required for SIC method and the performance is also restricted by the inaccurate channel estimation and large normalized Doppler frequency.

3) ICI reconstruction/cancellation combined with ICI self-cancellation: It can obtain advantages for the above two categories. However, if the redundancy is not large enough, the ICI self-

cancellation is not significant. Besides, performance is also restricted by the inaccurate channel estimation and large normalized Doppler frequency.

Conventional Doppler effect cancellation methods e.g. ICI self-cancellation, ICI reconstruction and cancellation and ICI reconstruction/cancellation combined with ICI self-cancellation will need heavy computation complexity or large redundancy in the receiver, and the performance will also be restricted by inaccurate channel estimation and large normalized Doppler frequency.

In case of ongoing 3GPP Release-14 work item related to vehicle-to-vehicle communication [3GPP16-RP152293], for example carrier frequency is 6 GHz and the vehicle speed is 140km/h (the maximum absolute vehicle speed may be 280 km/h), the normalized Doppler frequency will be up to 10.3%. Demodulation reference signal (DMRS) can be used to estimate channel in physical sidelink share channel (PSSCH)/ physical sidelink control channel (PSCCH). Based on the assumption that channel varies linearly during two neighboring DMRS sequences (i.e. DMRS symbols), DMRS symbols can be used to effectively estimate channel by using linear interpolation method. However, when channel varies fast and non-linearly during two neighboring DMRS symbols, the accuracy of channel estimation will be significantly degraded.

In current DMRS in legacy LTE, the separation of two neighboring DMRS symbols is one slot, i.e. 0.5 ms. A total of 7 symbols represent one slot in case normal cyclic prefix is used. On the other hand, in extended cyclic prefix, one slot is a combination of 6 symbols. In the high Doppler case for V2V, for example carrier frequency is 6GHz and the vehicle speed is 140 km/h (the maximum absolute vehicle speed will be 280 km/h), i.e. δ is about 0.1, the number of the channel average fading cycle is about 0.7 and 0.6 during two neighboring DMRS symbols for the normal cyclic prefix case and the extended cyclic prefix case, respectively. Therefore, the channel varies non-linearly during two neighboring DMRS symbols for current DMRS and the current DMRS is not enough in the high Doppler case for V2V. In [3GPP16-36885], working assumption is to increase DMRS density to 4 symbols per 1 ms with reusing PUSCH DMRS sequence in each physical sidelink channel except for PSBCH. An Illustration of increasing DMRS density (for example, 4 DMRS symbols per subframe) from legacy LTE (2 DMRS symbols per subframe) is shown in Figure A4-2.

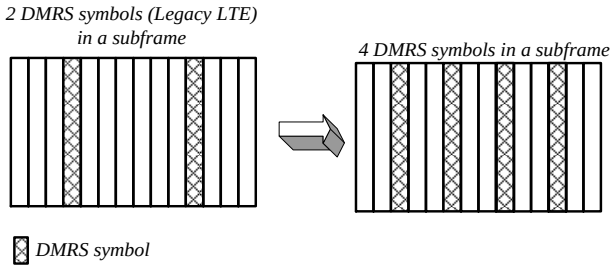
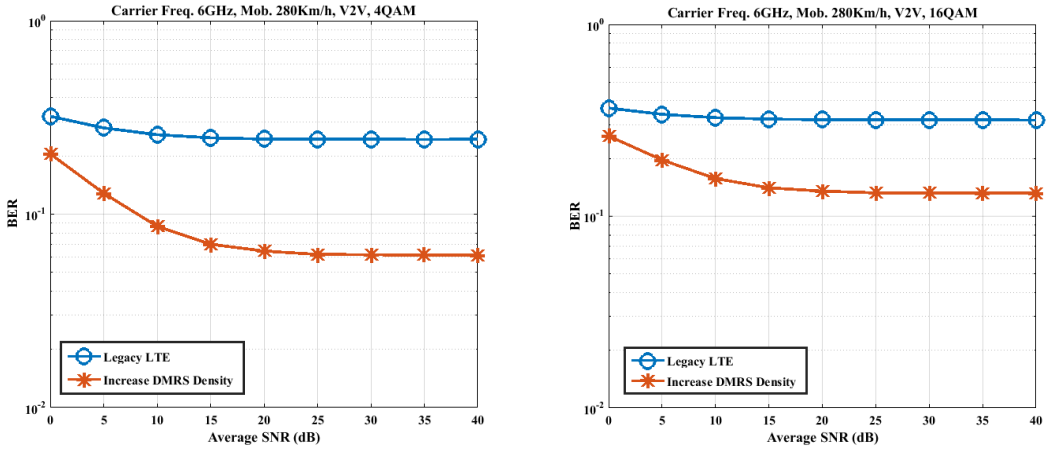


Figure A4-2: An Illustration of increasing DMRS density from legacy LTE



(a) (b)

Figure A4-3: Performance comparisons between legacy LTE and increasing DMRS density for (a) 4QAM and (b) 16QAM in the high Doppler case

Performance comparisons between legacy LTE (2 DMRS symbols per subframe) and increasing DMRS density (4 DMRS symbols per subframe) for 4 QAM and 16QAM in the high Doppler case for V2V are shown in Figure A4-3(a) and Figure A4-3(b), respectively. Spatial channel model (SCM) based on 3GPP TR 25.996 is used in the simulation. And the BER performance is the bit error rate after demodulation and before channel decoding. When the BER is larger than 10^{-1}, the system performance is not good enough even we further use channel coding system. Besides, channel is estimated by DMRS through linear interpolation, and no advanced ICI cancellation design is used in the receiver for this simulation. From the simulation results, we can observe that the performance of the legacy LTE is not good enough due to inaccurate channel estimation and large ICI for both 4 QAM and 16 QAM. Moreover, the method with 4 DMRS symbols per subframe may be good enough in 4 QAM but not good enough in 16 QAM, this is because the system performance is more sensitive to ICI in 16 QAM modulation. Therefore, Increasing DMRS density is not enough for high order modulation, for example 16 QAM, in the high Doppler case if advanced ICI cancellation and channel estimation methods are not adopted in the receiver.

A.5 State transitions between Connected and Connected Inactive

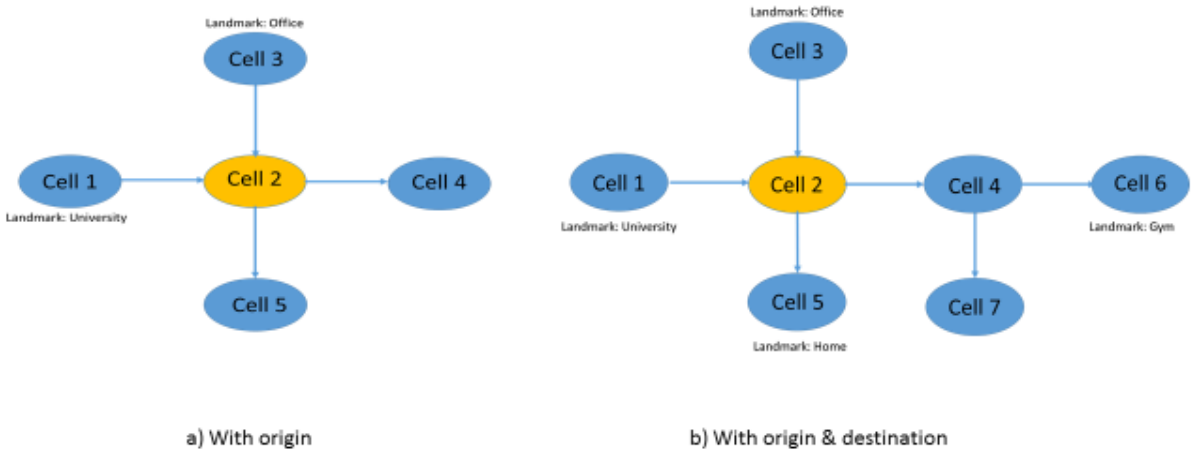


Figure A5-1: Markov model with additional context

Markov based prediction can be extended by using information about origin of the user as shown in example figure A5-1a. Here prediction at cell 2 will consider user statistics only when originated from a specific landmark instead of considering entire user statistics. The probability of transition is obtained as,

$$P(BS_n \longrightarrow BS_{n+1}/Origin) = \frac{N(BS_n \longrightarrow BS_{n+1}/Origin)}{N(BS_n/Origin)}$$

Where $BS_n \longrightarrow BS_{n+1}$ indicates transition of user from cell n to cell n+1, $N(BS_n \longrightarrow BS_{n+1})$ indicates number of times a user in cell n transited to cell n+1, $N(BS_n)$ indicates number of times user was found in cell n.

The model can be further extended using information of both origin and destination as shown in example figure A5-1b. Here prediction at cell 2 will consider user statistics only when originated from a specific landmark and travelling to a specific destination. The probability of transition is obtained as,

$$P(BS_n \longrightarrow BS_{n+1}/Origin\&Destination) = \frac{N(BS_n \longrightarrow BS_{n+1}/Origin\&Destination)}{N(BS_n/Origin\&Destination)}$$

Similar to next cell prediction, prediction of next route R_{n+1} can be done at each crossroad C_n (as indicated in figure A5-2) ,

$$P(R_n \longrightarrow R_{n+1}) = \frac{N(R_n \longrightarrow R_{n+1})}{N(R_n)}$$

$$P(R_n \longrightarrow R_{n+1}/Origin) = \frac{N(R_n \longrightarrow R_{n+1}/Origin)}{N(R_n/Origin)}$$

$$P(R_n \longrightarrow R_{n+1}/Origin\&Destination) = \frac{N(R_n \longrightarrow R_{n+1}/Origin\&Destination)}{N(R_n/Origin\&Destination)}$$

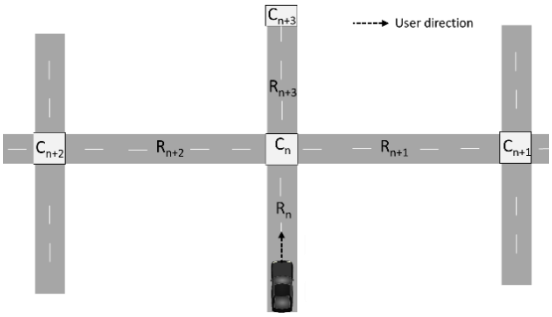


Figure A5-2: Route prediction

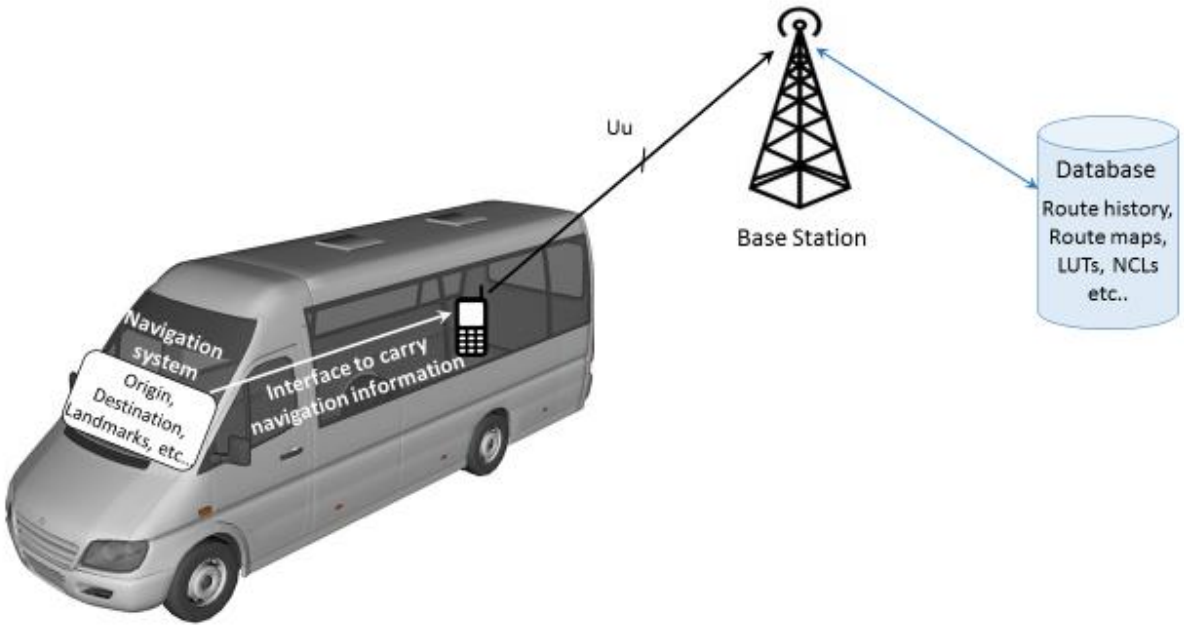


Figure A5-3: Interface in vehicular infrastructure

Figure A5-3 shows the new interface required in vehicular infrastructure to obtain the information about origin, destination, perfect route information etc and relay this context information to serving base station.

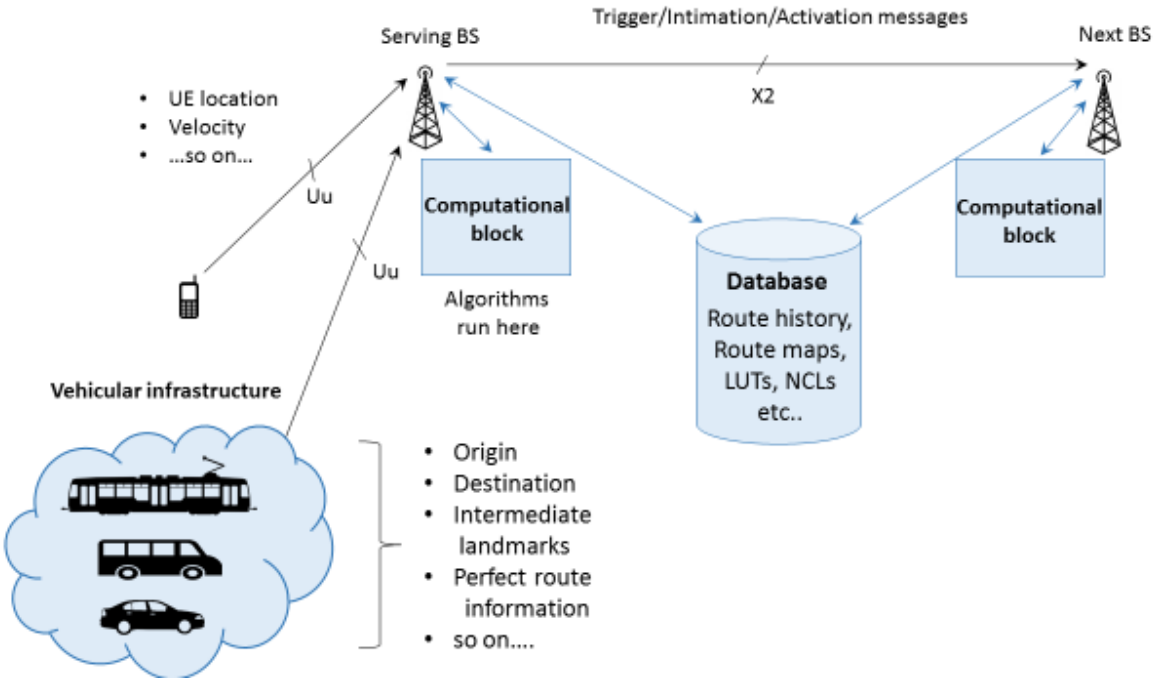


Figure A5-4: Overall architecture

Figure A5-4 shows the overall architecture to enable mobility context awareness. The information such as location, velocity etc., need to be signaled from UE to the serving base station via Uu interface. The information on user origin, destination, perfect route information etc., extracted from vehicular infrastructure need to be signaled to serving base station. The context message (trigger or intimation) needs to be conveyed from serving BS to predicted next BS to enable resource reservation or management.