Project**:** H2020-ICT-2014-2 5G NORMA

Project Name:

5G Novel Radio Multiservice adaptive network Architecture (5G NORMA)

Deliverable D4.1
RAN architecture components – intermediate report

Date of delivery:	30/11/2016	Version:	1.0
Start date of Project:	01/07/2015	Duration:	30 months

Document properties:

Document Number:	H2020-ICT-2014-2 5G NORMA/D4.1
Document Title:	RAN architecture components – intermediate report
Editor(s):	Sina Khatibi, Mark Doll
Authors:	Danish Aziz, Hajo Bakker, Mark Doll, Diomidis Michalopoulos, Vinh Van Phan, Peter Rost (Nokia); Vincenzo Sciancalepore (NEC); Jorge Rivas (Atos); Ignacio Berberana (Telefonica I+D); Alessandro Colazzo (Azcom); Sina Khatibi (Nomor); Vasilis Friderikos, Oliver Holland (King's College London); Bin Han, Shreya Tayade (TU Kaiserslautern); Dario Bega (Universidad Carlos III de Madrid)
Contractual Date of Delivery:	30/11/2016
Dissemination level:	PU[1]
Status:	Final
Version:	1.0
File Name:	5G NORMA D4.1 Part I

Revision History

Revision	Date	Issued by	Description
1.0	30.11.2016	5G NORMA WP4	Version delivered to project officer.

Abstract

The main goal of 5G NORMA is to propose a multi-service mobile network architecture that adapts the use of the mobile network resources to the service requirements, the variations of the traffic demands over time and location, and the network topology. This is the first public deliverable of WP4 containing the findings of WP4 after the second 5G NORMA design iteration. In the current design the functional blocks are categorised into four groups of data layer, distributed control, common control, and dedicated control covering both LTE and 5G NORMA innovations for 5G. Three options for network slicing (slice-specific protocol stacks up to including upper PHY, full sharing with dedicated RRC and common MAC) and three solutions for multi-connectivity (common TCP/IP, common PDCP, and common MAC) are discussed. Finally, background, the 15 individual 5G NORMA innovations that form the basis of the above results, are presented in detail and with further references in the second part of this deliverable.

Keywords

5G, network architecture, RAN architecture components, functional allocation and (de-)composition, multi-connectivity, multi-tenancy, network function virtualization, network slicing

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

PU = Public

Executive Summary

This document provides a detailed view of the radio access network (RAN) architecture components associated with the 5G NORMA project. In particular, after providing an overview of the relevant state of the art, an extensive analysis of the decompositions of network functions is presented, aligned with the concept of multiservice adaptive network architecture.

Within this context, network functions corresponding to different layers of the protocol stack are scrutinized on the basis of their operation towards a flexible network design. The considered network functions are first sorted according to the layer of the protocol stack they belong to, followed by a description of their operation as well as their technical characteristics. Then, the above network functions are further grouped into *function blocks*: A function block is a set of network functions with similar technical operation, thereby associated with a similar role in the network.

Capitalizing on the aforementioned classification of network functions into function blocks, this document focuses on listing the detailed requirements of the function blocks as well as their interdependencies with each other. To this end, along with a brief description and a detailed sketch of their operation, a list of interfaces is provided for each of the considered function blocks. In this regard, their utility as operational elements are portrayed through detailed functional sketches, while their roles as part of the network are highlighted via their dependency with other function blocks as well as other network elements.

The resulting 5G NORMA control and data layer functional architecture, is based on the current 3GPP LTE RAN architecture to equally support LTE and novel 5G RATs. For 5G, the data layer is modified and extended specifically at MAC and PDCP level for integration of multi-connectivity and multi-RAT support. The control layer has been reworked to add network programmability and RAN slicing. Most control functions run as SDMC-enabled control applications on top of the SDM controller (SDM-C) for slice-dedicated and the SDM coordinator (SDM-X) for shared functionality. Only time critical and computational expensive radio resource control (RRC) and management (RRM) functionalities are kept as “legacy” distributed network functions. Control layer functionality has been extended for multi-tenancy, multi-service, multi-connectivity and multi-RAT support.

The concept of multi-connectivity across heterogeneous networks represents one of the focal points of 5G NORMA. This document presents the multi-connectivity techniques proposed within the context of this project, which span from heterogeneous approaches within the same radio access technology to multi-connectivity mechanisms across different radio access technologies. A list of the involved network functions is provided in detail, including both the functions already existing in LTE and needed to be modified, and the new functions required for 5G. Such network functions are designed in a flexible fashion, in the sense that they can be smartly placed within the network and thus be consistent with the 5G NORMA concept of flexible allocation of network functions.

A considerable part of this document is dedicated to the innovations proposed within the 5G NORMA framework. Such innovations come along with novel architectural developments on the RAN, therefore they are conceptually highly related to the network function decomposition discussed above. In order to provide an intelligible view of the 5G NORMA approach, the innovations developed in this project are introduced twice: First, they are listed in a concise format where their connections to the aforementioned function blocks are highlighted. In this respect, the relationships of the 5G NORMA novelties with the concept of network function decomposition are underlined, pinpointing thus the significance of this approach for achieving the desired network requirements. Then, an extensive description of the 5G NORMA innovations is presented in the document’s second part, along with a detailed view on how such innovations fit within the notion of 5G.

Table of Contents

List of Figures

List of Tables

List of Acronyms and Abbreviations

Term	Description
3GPP	3rd Generation Partnership Project
A/D	Analog-to-Digital Converter
ABSF	Almost-Blank Sub-Frame
AI	Air Interface
ANDSF	Access Network Discovery and Selection function
AP	Access Point
ARP	Allocation and Retention Priority
ARQ	Automatic Repeat Request
CAPEX	Capital Expenditure
CCMCC	Coordinated and Cooperative Multi-Connectivity Cluster
CD	Confidential Degree
CID	Context Identification
CN	Core Network
CoMP	Coordinated Multipoint Transmission and Reception
CP	Control Plane
CPRI	Common Public Radio Interface
CQI	Channel Quality Indication
CRC	Cyclic Redundancy Check
CRE	Cell Range Expansion
CRS	Common Reference Signal
D/A	Digital-to-Analog Converter
D2D	Device-to-Device
DCN	Data Communication Network
DL	Downlink
DPF	Direct Provisioning Function
DRX	Discontinuous Reception
E2E	End-to-End
EC	European Commission
eIMTA	enhanced Interference Mitigation and Traffic Adaptation
eMBSFN	Evolved Multimedia Broadcast Single Frequency network
eMBMS	Evolved Multimedia Broadcast Multicast Services
EPC	Evolved Packet Core
ETSI	European Telecommunication Standard Institute
E-UTRAN	Evolved Universal Terrestrial Radio Access Network
FDD	Frequency Division Duplexing
FFT	Fast Fourier Transform
FE	Function Element

FEC	Forward Error Correction
GBR	Guaranteed Bit Rate
GDB	Geolocation Database
GTP	GPRS Tunnelling Protocol
GWCN	Gateway Core Network
H2020	Horizon 2020
HARQ	Hybrid Automatic Repeat Request
HetNet	Heterogeneous Network
HO	Handover
HSPA	High Speed Packet Access
ICIC	Inter-Cell Interference Cancelation
ICT	Information and Communication Technologies
IoT	Internet of Things
OFDMA	Orthogonal Frequency-Division Multiple Access
LNA	Low-Noise Amplifier
LSA	Licensed Shared Access
LTE	Long-Term Evolution
MAC	Medium Access Control
MBSFN	Multimedia Broadcast Single Frequency network
MBMS	Multimedia Broadcast Multicast Services
MBR	Maximum Bit Rate
MCH	Multicast Channel
MCN	Multi-hop Cellular Network
MCPTT	Mission Critical Push to Talk
MEC	Mobile Edge Computing
MIB	Master Information Block
MIMO	Multiple-Input Multiple-Output
MME	Mobility Management Entity
MMC	Massive Machine Communication
MMSE	Minimum Mean Square Error
MN	Moving Network
MNO	Mobile Network Operator
MOCN	Multi-Operator Core Network
MPTCP	Multi-Path Transmission Control Protocol
MTA	Moving Tracking Area
MTC	Machine Type Communication
MTCH	Multicast Traffic Channel
MUX	Multiplexer
MVNO	Mobile Virtual Network Operators

NAS	Non-Access Stratum
NFV	Network Function Virtualisation
NVS	Network Virtualisation Substrate
OPEX	Operational Expenditure
PA	Power Amplifier
PDCP	Packet Data Convergence Protocol
PDU	Protocol Data Unit
PHY	Physical Layer
ProSe	Proximity Services
PS	Public Safety
PSM	Power Saving Mode
QCI	QoS Class Indicator
QoE	Quality of Experience
QoS	Quality of Service
QPS	QoS Parameter Set
RAN	Radio Access Network
RAT	Radio Access Technology
RB	Radio Bearer
RLC	Radio Link Control
ROHC	Robust Header Compression
RRC	Radio Resource Control
RRH	Remote Radio Head
RRM	Radio Resource Management
RRU	Radio Resource Utilisation
RTT	Round Trip Time
SAE	System Architecture Evolution
SCTP	Stream Control Transmission Protocol
SDMC	Software Defined Mobile Network Control
SDMN	Software Defined Mobile Network
SDU	Service Data Packet
SF	Service Flow
SFN	System Frame Number
S-GW	Service Gateway
SLA	Service Level Agreement
SON	Self-Organised Network
SPS	semi-persistent scheduling
SIB	System Information Block
TB	Transmission Block
TCP	Transmission Control Protocol

TDD	Time Division Duplex
TTI	Transmission Time Interval
TVWS	TV White Spaces
UCA	User-Centric Connection Area
UDN	Ultra-Dense Network
UE	User Equipment
UMTS	Universal Mobile Telecommunications System
UL	Uplink
UP	User Plane
URC	Ultra-Reliable Communication
V2X	Vehicular-to-everything
VNF	Virtual Network Function
VoIP	Voice over IP
WP	Work Package

Definitions

The key terms and definitions in this report are as follows:

- **Almost-Blank Sub-Frame (ABSF):** are sub-frames without scheduled data transmissions, thus reducing interference to neighbouring cells during those sub-frames at the cost of fewer sub-frames available for scheduling data transmissions.
- **Central Cloud:** The central cloud comprises one or more centrally located data centers hosting a significantly large collection of processing, storage, networking, and other fundamental computing resources. Typically, only a few of them are found in a nationwide operator network.
- **Core Network (CN)**: Refers to network elements and the functions required to offer numerous services to the customers who are interconnected by the access network. In the 3GPP EPS this is the EPC, which handles all NAS functions.
- **Edge Cloud**: The edge cloud comprises a small, locally located, i.e. close to or at the radio site, collection of processing, storage, networking, and other fundamental computing resources. Typically, the number of edge clouds is at least one order of magnitude higher than the number of central cloud instances.
- **Enhanced ICIC (eICIC)**: Refers to the method used in LTE to manage the interference using ABSF.
- **Evolved Packet Core (EPC):** Evolved Packet Core is the core network part of the Evolved Packet System (EPS) system. It serves as the equivalent of GPRS networks.
- **GPRS Tunnelling Protocol for Control plane (GTP-C):** It is used within the GPRS core network for signalling between gateway GPRS support nodes (GGRS) and serving GPRS support nodes (SGSN). This allows the SGSN to activate a session on a user's behalf, to deactivate the same session, to adjust QoS parameters, or to update a session for a subscriber who has just arrived from another SGSN.
- **Home Subscriber Server (HSS):** The Home Subscriber Server is a database that contains user-related and subscriber-related information. It also provides support functions in mobility management, call and session setup, user authentication and access authorisation.
- **Inter-Cell Interference Coordination (ICIC)**: Refers to the method used in LTE to manage the interference arising due to signal coming from adjacent cell sites. The basic principle is to coordinate the scheduling of cell edge users in a way that users are not scheduled on the same frequency time resources as users in other cell (dynamic fractional frequency reuse).
- **Mobility Management Entity (MME):** The MME is the main signalling node in the EPC. It is responsible for initiating paging and authentication of the mobile device. It retains location information at the tracking area level for each user and then selects the appropriate GW during the initial registration process. It also plays a vital part in handover signalling between LTE and 2G/3G networks.
- **Multi-tenancy:** Offering an own virtualised logical network to a tenant who has exclusive and secure access to its own network's virtual resources.
- **Multi-connectivity**: Refers to the scenario, where the UE is connected to two or more radio access points at the same time.
- **Network Function Virtualisation (NFV)**: Refers to a network architecture philosophy that utilises virtualisation technologies to manage core networking functions via software as opposed to having to rely on hardware to handle these functions. The NFV concept is based on building blocks of virtualised network functions that can be combined to create full-scale networking communication services.
- **Packet Data Network Gateway (PDN-GW):** The Packet Data Network Gateway is the point of interconnect between the EPC and the external IP networks.

- **Radio Access Network (RAN)**: Subsumes the network elements and the functions involved that connect individual devices to other parts of the network through radio connections. In the 3GPP EPS this is the E-UTRAN, which handles all access system functions.
- **Radio Access Technology (RAT)**: The interface technology employed by the RAN. Examples are Universal Mobile Telecommunications System (UMTS, i.e. 3GPP UTRA); Long Term Evolution (LTE, i.e. 3GPP E-UTRA); WiFi; 5G.
- **RAT agnostic:** it could imply the following:
    - The same implementation but different parameterisation is used;
    - The same processing instance is used for two different RATs.
- **Serving Gateway (S-GW):** Serving Gateway routes and forwards user data packets between the UE and the external networks. It's the point of interconnect of the EPC to the RAN (E-UTRAN).
- **Software Defined Networking (SDN):** Refers to an approach to computer networking that allows network administrators to manage network services through abstraction of higher-level functionality. This is done by decoupling the control layer from the data layer.
- **Software Defined Mobile Network Controller (SDM-C)**: The SDMC is responsible for managing resources allocated on each network slice and controlling the functions executed in the edge and network clouds.
- **Software Defined Mobile Network Orchestrator (SDM-O):** Responsible for allocating resources according to network slice requirements. Management of the inter-slice lifecycle.

PART I: ARCHITECTURE

1. Overview

1.1. Objective of the Document

This is the first public deliverable of WP4 containing the findings of WP4 after the second 5G NORMA design iteration. The document is divided into two parts. The first part presents the outcome of joint effort of WP4 partners to derive a common control and data layer functional architecture capable of supporting all specific partner innovations. The presented functional architecture aims at implementing the three 5G NORMA innovative enablers, namely adaptive (de)composition and allocation of mobile network functions, joint optimisation of mobile access and core network functions, and Software-Defined Mobile Network Control (SDMC). Particularly the latter involves joint efforts of both WP4 and WP5, and it will be a focus in the following harmonization and integration within WP3, which will be document in D3.2. This deliverable puts together all partners' proposals for the two central functionalities network slicing and multi-connectivity, providing a conclusive set of alternatives for these two crucial functionalities. It is noted that such functionalities are needed in order to enable 5G NORMA's innovative capabilities of multi-service and context-aware adaptation of network functions, as well as mobile network multi-tenancy.

In the second part of this report, all specific partner innovations are detailed and put into the context of the functional architecture.

1.2. Structure of the Document

After a brief overview in Section 1.3 on the innovative functionalities and scientific highlights of this report, Section 2 presents a short summary of relevant state of the art, namely ETSI NFV and 3GPP LTE, which form the technical foundations of 5G NORMA's RAN design. In addition, work performed by the European projects METIS and iJOIN as well as the Small Cell Forum are discussed, which provide the basis for different functional RAN splits.

Section 3 presents the key architectural innovations for flexible network design. Section 3.1 introduces the overall functional architecture; Section 3.2 explains the interfaces among its SDMC-enabled function blocks; and Section 3.3 focuses on RAN Slicing. A detailed description of all function blocks of the presented architecture is provided in Annex Section 6.

Section 4 presents the concept of multi-technology architecture in heterogeneous networks by covering multi-connectivity in Section 4.1, radio resource and connection management in Section 4.2, and the concept of multi-tenant HetNets in Section 4.3. Regarding the multi-connectivity, the common TCP/IP, the common PDCP, and common MAC approach are explained and their respective applicability (Section 4.1.4) is discussed.

Section 5 concludes the first part of this report including an outlook to the final third design iteration. Annex Section 6 provides detailed information about the function blocks of the functional architecture presented in Section 3.

The second part of D4.1 comprehensively describes the innovations of each of the partners In detail, Section 7 discusses the individual novel functionalities, i.e.,

- User-Centric Connection Area,
- RAN support for optimised on-demand adaptive network functions and services,
- Mobile Edge Computing and Network Resource Allocation for Multi-Tenancy,
- Multi-Service Technologies,
- Multi-tenant dynamic resource allocation,
- Multi-RAT Integration,
- Data layer and Control layer Design for Multi-Connectivity,
- Flexible 5G service-flow (SF) with in-SF QoS differentiation and multi-connectivity,
- Multiple connectivity at the different layers,

- Centralised Radio Resource Management,
- Geolocation Databases, Use of Geolocation Information and Associated Opportunities,
- Functional (de)composition for Supporting Multi-Tenancy in HetNets,
- Multi-Tenancy in Multi-RAT Environments,
- Load Balancing of Signalling Traffic,
- QoS innovation.

Section 8 complements the second part of this report by discussing the integration of these novel functionalities into the functional architecture derived by WP4.

1.3. Novel 5G NORMA Functionality

This section provides a summary of the novel functionality developed within the framework of 5G NORMA WP4. For a better understanding of this novel functionality, we first provide an overview of the scientific highlights of the project, followed by a brief description for each of the specific technical innovations introduced by 5G NORMA.

Scientific highlights: The innovations proposed by 5G NORMA require architectural enhancements within the control and data layer of both radio access network (RAN) and core network (CN). In order to carry out such innovations and thereby accommodating them into the proposed architecture, the project adopts a series of solutions which mainly comprise the following four actions, i.e., a) proposal of new function blocks; b) partial modification of function blocks already existing in LTE systems; c) new interfaces which support either the interaction of function blocks or the flexible placement at different physical elements in the network, or both; d) incorporation of virtualisation techniques into the overall architecture.

The technical innovations introduced by 5G NORMA are listed below. Further details and background information can be found in D4.1 Part II.

Cloud-based RAN architecture: 5G NORMA supports the use of a new cloud-based RAN architecture that differs from the conventional distributed RAN architecture in LTE in the fact that the radio resource control (RRC) and Packet Data Convergence Protocol (PDCP) layers are located in a central location and shared among different access points. This cloud-based architecture facilitates the support of ultra-reliable applications, and it hides frequent mobility events from the central cloud. The new functionalities involved are as follows: Modified data transfer and routing at the PCDP level; novel flow control mechanisms; novel mapping between radio bearer service (RLC layer and below) and service flow (PDCP layer); user-layer control-layer separation; duplicate detection and reordering of RLC PDUs; and modified buffering of PDCP PDUs. The function blocks which are affected represent part of the RRC layer and are related to the user- and cell-oriented functionalities (they are defined in Section 3 as RRC User; RRC Cell function blocks). The PDCP layer is also affected, in particular the functionalities related to the steering of data packets (the corresponding function blocks are defined as PDCP, PCDP Split Bearer in Section 3). In addition, the support of ultra-reliable applications via the cloud-based architecture involves the modification of the following radio link control (RLC) function blocks as well, i.e., RLC TM; RLC AM, and RLC UM-AM.

Support of multi-tenancy: In order to extend the 3GPP framework for network sharing [31.130] with the objective of meeting the 5G multi-tenancy requirements, we are proposing two innovations. First, a new sharing criterion is introduced that allows for allocating the resources among tenants in a more flexible way. Our criterion takes as input: (i) a set of resources required from the different tenants, (ii) the corresponding period of time, (iii) their location, (iv) the quality-of-service (QoS) constraints, and (v) the probability of rejecting a call. Based on this input, the criterion is used to allocate the available resources in order to optimise and maximise the resources' utilisation. There are different possible optimisations that can be followed, e.g., maximizing the number of tenants or the overall profit. With the proposed criterion, resources are only reserved for tenants when they are needed and they are released when they are not used and available for

other tenants. Furthermore, the tenants do not need to ask (as in 3GPP) for more resources because our criterion provides each tenant with the requested resources.

A second innovation is a new sharing mechanism, which is signalling-based and with no human intervention. It enables a more efficient sharing of the network resources according to SLA requirements and taking into account also the commercial agreement. Given the amount of information involved (including the channel quality of each user) and its dynamic nature, the mechanism should be distributed. Since it may be triggered frequently (whenever a user joins, leaves or changes its location), it should also be computationally efficient. It should also control the number of handoffs triggered, as those may represent high signalling overhead. Two ways of implementing the mechanism are envisaged:

i. One new multiplexer function with a scheduler for each network slice;
ii. One scheduler for the network that allocates resources among different slices.

Therefore, we need to design completely new scheduling and multiplexing functions with new algorithms to enable new 5G functions for multi-tenancy, both in single-RAT and multi-RAT scenarios. They will impact the function blocks such as medium access control (MAC) Scheduling for supporting radio resource management (RRM) dynamic resource allocation and handovers.

MAC-level multi-connectivity: 5G NORMA investigates a MAC-level multi-connectivity (MC) solution, enabling RAN level supports for low-latency and high-reliability in small-cell ultra-dense networks (UDN). This MAC-level MC functionality is using broadcast based uplink transmission or single-frequency-network (SFN) based downlink transmissions between UE and a coordinated and cooperative MC cluster (CCMCC) of local small-cell access points (APs) in proximity of the UE (centric to UE), as configured and controlled by the serving RAN. This functionality has implications on RRC, MAC, optionally PHY and RLC for certain alternative solutions, and function blocks related to c-layer RAN level network interfaces (similar to LTE S1/X2).

Inter RAT multi-connectivity: Current dual connectivity functionalities in 3GPP LTE do not address the scenario of two base stations belonging to different radio access technologies (RATs). As 3GPP LTE is a widely accepted and heavily deployed technology, the transition from LTE to 5G is critical, and will take some time. Therefore, it is of high importance to consider backward compatibility of 5G with previous standards such as LTE. In 5G NORMA, we propose a multi-connectivity scheme that enables a UE to connect to multiple RATs simultaneously. Based on the data traffic, one of two different operating modes can be selected: reliability mode or diversity mode. Reliability mode allows for transferring the same data packets across multiple links, i.e., data duplication, thereby ensuring a reliable data transfer. Diversity mode allows for transferring control and data layer data across either one of the connected RATs, increasing the number of users that can be simultaneously supported. To enable dynamic MC of different RATs to a UE, new functions such as inter-RAT packet scheduling and inter-RAT link scheduling will be designed. The impacted function blocks are PDCP, PDCP Split Bearer, and MAC Scheduling (RRM).

Optimised control of radio multi-connectivity: 5G NORMA considers flexible context-aware adaptive use of master-slave RRC connections for facilitating optimised control of RAN level MC, especially when multiple RATs are involved or a UE being served requires ultra-high reliability or low latency. This functionality includes on-demand dynamic setup and release of slave RRCs as well as optimised delegation or distribution of RRC control functions and procedures for slave RRCs. This functionality has direct implications on function blocks related to RRC and c-layer RAN level network interfaces (similar to LTE S1/X2).

User centric connection area: 3GPP LTE is mainly designed for the transmission of broadband packet payloads. Due to this, signalling mechanisms are inefficient for small packet payloads. 5G NORMA proposes a framework of a User-centric Connection Area (UCA), which reduces the signalling load on the air interface and towards the mobility management (e.g., MME in 3GPP LTE). A UCA is defined as a coverage area where the UE-context is known in advance to all 5G APs. For each UE with small or sporadic data, an individual UCA is configured by the RAN, e.g.

an anchor node based on SON functionality (neighbour relation table) and with certain assistance of the UE (measurement of surrounding APs).

A UE, even if no data has to be transmitted, permanently remains in the connected state and will be transferred to a sub-state defined as UCA_enabled. Within this sub-state, the UE will not report any channel measurements (channel quality information (CQI)) or other measurements, i.e., it will perform cell reselection without control of the access network (forward-handover). With the help of open loop synchronisation and efficient access protocols for 5G as described in [SWS15], the UE is able to perform both contention based and contention free UL (uplink) transmissions in any cell (measured as best server by the UE) within the UCA. This UL transmission may carry the notification of the best server or the small uplink user data. The best serving node will forward these UL packets to the anchor node. In the case of DL (Downlink) transmission, the anchor node (e.g. with the help of best server updates) will forward the packets to the currently best serving node of the user. If the serving node is not known, the anchor node will trigger a paging message within the area of the UCA to identify the best serving node of the UE. If the UE leaves the coverage of the UCA, a dedicated signalling (compared to the LTE handover procedure) will define a new user specific UCA. The UCA framework minimises the RAN and CN signalling overhead related to connection management (idle/active transitions) and mobility management (paging, handover).

The UCA framework has impact on the following function blocks: PHY Transmission Point (TP), MAC Scheduling (RRM), RRC, Self-Organizing Networks (SON), RAN Paging and Transport (SDN).

Multi-connectivity of mm-wave access points: Future deployments of mm-wave access points (mmAPs) in 5G access networks will ensure the delivery of high data rates to the UEs. However, it is challenging to provide highly reliable and uninterrupted data transfer to the UEs using the mm-wave technology (especially for the mobile UEs). The urban-micro mm-wave channel, as considered for the 5G NORMA architecture, is characterised by a low number of possible paths (LOS and NLOS) between base station and UE, from which in most cases only one path will be used for transmission with high gain, narrow half-power beam width antenna beams. This makes transmission quality sensitive to blocking effects caused by sudden user movement or obstacles entering the transmission path, leading to poor reliability. Therefore, to minimise interruption times or ideally even to avoid interruptions and to guarantee reliability, 5G NORMA proposes that the mmAP deployments must be supported by the low-band 5G coverage and a redundant coverage of mmAPs should be provisioned for the UEs.

For this purpose, multi-connectivity (MC) will be an essential or rather a fundamental feature in 5G-access networks. Moreover, a UE must be able to detect and receive multiple mmAPs to ensure the possibility of MC, link monitoring, and fast selection. To provide redundant coverage, multiple mmAPs are placed within the low-band 5G coverage area, building a "serving cluster", so that the UEs are within transmission range of each mmAP of the cluster. It is assumed that a UE is served by at least one of the mmAPs out of the serving cluster at a time. If the connection to the serving mmAP is blocked by an obstacle, the UE possibly will be instructed to connect to another mmAP serving the area from another direction, so that the transmission is no longer affected by the obstacle, i.e., there is no interruption in the data transfer as another mmAP can take over. It is assumed that such a cluster of mmAPs is within the coverage area of a 5G eNodeB, and the mmAPs are using the same high carrier frequency and bandwidth. However, for full flexibility, the mmAPs in a cluster may belong to different 5G eNodeBs. An mmAP-based MC has impact on the following function blocks: PHY TP (new mm-wave air interface), RRC, PDCP, PDCP Split Bearer and SON.

(De-)centralised radio resource management: The 5G NORMA architecture provides a RAN platform for flexible decomposition and allocation of RAN functionalities. This architecture enables radio resource allocation in a centralised and distributed manner, i.e., resources can be scheduled dynamically regarding UL/DL load capacity conditions. The scheduling takes into account the service requirements (e.g., data rate and latency) as well as deployment characteristics

(e.g., frontal/backhaul latency) in order to select optimally the RRM functional distributions within the network. This framework will impact MAC Scheduling (RRM), RRC Cell and RRC User function blocks (responsible for TDD configuration).

SON based flexible configuration of 5G RAN protocol stacks: 5G NORMA enables self-organized networks (SON) based flexible (re-)configuration and (re-)distribution of the 5G RAN protocol stack including physical and virtual network functions (NFs). It enables the dynamic management of exposed capabilities of flexible 5G RAN including not only UE capabilities but also generic hardware/software and front-haul capabilities of individual APs. That is, the actual protocol stack structure and operation mode of an AP, not just cell specific system parameters or UE specific bearer-configuration or link adaptation parameters, may be dynamically changed or self-reconfigured on the fly to best serve current local users and services. This is in line with the overall concept of flexible network function decomposition and cloud-based placement and relocation of necessary function blocks in 5G NORMA. This functionality has implications on RRC, SON and operations and management (O&M) related function blocks.

UE agent based decomposition of E2E connection: 5G NORMA introduces a so-called UE agent based decomposition of E2E connections as part of the RAN support for providing optimised on-demand adaptive NFs and services to end users. This functionality decomposes or splits an E2E connection of the UE into two E2E connections: the first is between the UE and the UE agent located in the RAN or inside a RAN proxy server, and the second is between the UE agent and the remote E2E server. This enables a maximised utilisation of 5G RAN capability and capacity in terms of ultra-high data rate, reliability, and adaptability. It facilitates providing long-duration remote access services such as Internet or new 5G services across one or more different network domains, where one domain may cause a bottleneck effect in terms of capacity and adaptability for E2E connections. Examples of those network domains include much slower legacy wired or wireless networks, or satellite communications links. The bottleneck effect may be also caused by a data rate limitation on application level at either source or sink of the E2E connection. Using this functionality for uploading a large video clip for example, the UE may first send the clip to the UE agent over the first ultra-fast 5G connection within seconds and let the UE agent handle the upload to the remote server. This significantly improves energy efficiency and quality of experience (QoE) for the UEs. This functionality has implications on RRC User, PDCP, MAC Scheduling (RRM), non-access stratum (NAS), and SON function blocks.

RAN support for advanced QoS control: 5G NORMA considers RAN support for advanced QoS control, aiming for enhancing QoE for end users. This includes RAN level in-bearer QoS differentiation with application-aware in-bearer sub-flow based packet filtering and prioritizing, fast in-band d-layer control of in-bearer sub-flows, and possible UE assistance for optimizing related RAN functions and procedures. This functionality has implications on PDCP, MAC Scheduling (RRM), and RRC User function blocks.

QoS innovation: 5G NORMA designs as a multi-service adaptive mobile network architecture where network resources fulfil the service requirements in a flexible and dynamic way. This network concept will introduce new objectives and advanced QoS requirements that are important to consider and to evaluate. In 3GPP LTE, the QoS is related to the bearer model and uses a static set of QoS parameters associated with a QCI value that specifies the treatment of IP packets received on a specific bearer. 5G NORMA proposes new dynamic methods and parameters in order to satisfy the upcoming requirements of new services.

The proposal aims to break the QCI concept used in LTE, and to design and to implement a QoS method that provides flexibility to fit the QoS parameters to service needs and network resources. The intention is to define a set of QoS parameters (QPS) applicable at different network levels and dynamically configurable by different network controllers. It is also important to note that there are two different ways to define an e2e bearer depending on the service type, i.e., dynamic and static. Using the dynamic way, the QPS is set up on the fly, providing more flexibility. On the other hand, using the static way, the QPS is managed beforehand, achieving therefore more scalability. The function blocks that are impacted are MAC Scheduling (RRM) and QoS Control.

Geolocation database and geolocation-based management: 5G NORMA considers the use of geolocation capabilities for 5G applications, enabled by a "geolocation database" functionality. Although termed a "database" in various fields of operation hence that name being maintained, such a capability is also fundamentally responsible for taking management decisions using the geolocation information, also using knowledge about the environment, technical capabilities of elements, and various modelling approaches. This capability has traditionally arisen in spectrum management realms for the purpose of facilitating dynamic or opportunistic spectrum access, e.g., TV white space, or, more loosely, to assist spectrum sharing between operators through concepts such as Licensed Shared Access (LSA). However, while maintaining this as a separate function block in order to ensure compatibility with the requirements of some such applications, 5G NORMA expands the concept significantly to be compatible with numerous other purposes in the context of virtualised RANs. These include, for example, (i) the use of geolocation information to optimally place virtualised RAN elements in order to minimise propagation delay for ultra-low latency applications, (ii) the use of geolocation information to be able to better dynamically serve the spatially and temporally varying traffic requirements using virtualised RANs, and (iii) the use of geolocation information to assist rendezvous, particularly in highly heterogeneous networking scenarios.

2. State of the Art

Research in methods for optimally locating RAN functions is currently an open topic that will experience a push due to the rise of NFV. The European Telecommunications Standards Institute (ETSI) has defined an NFV architecture that is a key enabler for the flexible function (de)composition and allocation concept of 5G NORMA.

Initiatives exist that have focused on on-demand RAN function decomposition leveraging the benefits of NFV to enable a cloud-enabled base station, where RAN functionality has been flexibly centralised through an open IT platform based on a cloud infrastructure [IJOIN-D51].

Another aspect addressed in literature is, for example, joint RAN and backhaul optimisation methods. There are studies, e.g., [PMD+15], focused on the design of novel MAC and RRM schemes for constrained, non-ideal backhaul, which can influence the RAN performance.

5G NORMA follows the same NFV based function decomposition and aims to enhance it considering the function decomposition depending also on the service type and considering its objective QoS requirements.

2.1. Network Function Virtualisation

3GPP has adopted ETSI NFV MANO [NFV-002] shedding light on the potential impact of virtualised networks on the existing 3GPP SA5 network management architecture [32.842], considering either partial or entire adoption of VNFs with respect to macro-base stations and core network elements. The objective is to identify requirements, interfaces, and procedures, which can be re-used or extended for managing virtualised networks. In Release 14, 3GPP has introduced a specification on architecture requirements for virtualised network management [28.500], considering complementary specifications on configuration, fault performance, and life-cycle management. An equivalent study focusing on small cells and on the adoption of flexibly centralised RANs is performed by the Small Cell Forum [SCF-15b].

Additionally, in an effort to support devices and customers with different service characteristics, including vertical market players, 3GPP SA2 has introduced the support for separate data communication networks (DCNs) in Release 13 [23.401], with different operation features, traffic characteristics, and policies. Each DCN is assigned to serve a different type of users based on subscription information. It assures resource isolation and independent scaling, offering specific services and NFs including RATs [SPS+16]. Effectively, the 5G network slice broker may allocate a collection of shared network resources and VNFs among particular slices that fulfil the requirements of certain communication services.

Finally, many evolving 5G services are envisioned to be offered closer to the user, in the network edge, in order to reduce latency and enhance perceived performance, e.g., by adopting the ETSI MEC paradigm [ETSI-MEC]. Flexible service chaining needs to be enhanced to establish dynamic services considering edge network locations, which may potentially be combined with VNFs in order to ensure a joint optimisation of services and networking. Edge server locations can also be exploited for storage, computation, and dynamic service creation by verticals and over the top (OTT) providers, introducing in this way another multi-tenancy dimension.

2.2. 3GPP LTE Functional Architecture

Mobile broadband services were already provided by 3G systems (WCDMA), e.g., HSDPA and HSUPA. Based on requirements such as high peak data rates, short round trip times as well as flexibility in frequency and bandwidth, the 3rd Generation Partnership Project (3GPP) started work on 4G with two major work items:

- Long Term Evolution (LTE), and,
- System Architecture Evolution (SAE).

These resulted in the specification of the Evolved Universal Terrestrial Radio Access (E-UTRA), the Evolved Universal Terrestrial Radio Access Network (E-UTRAN), and the Evolved Packet Core (EPC) within the 3GPP Release 8.

LTE utilises Orthogonal Frequency-Division Multiple Access (OFDMA) for DL transmissions and Single-Carrier Frequency-Division Multiple Access (SC-FDMA) for UL transmissions. OFDM based multiple access schemes divide the channel into multiple sub-carriers, and modulate the information on the sub-carriers. These sub-carriers are sent over the air as parallel data streams. LTE also leverages Multiple-Input Multiple-Output (MIMO) technologies to achieve high data rates on the air-link, e.g., both 2x2 and 4x4 MIMO options are available [Tech-Lib].

To fulfil the above mentioned requirements, to speed up the connection set-up, to improve hand-over execution time, and to enable a faster scheduling, it was decided to omit the 3G radio network controller (RNC) node but to move RNC functions to APs, denoted as evolved NodeBs (eNodeBs). This enables a flat E-UTRAN architecture, consisting only of eNodeBs, which provide the data layer (PDCP/RLC/MAC/PHY) and control layer (RRC) protocol terminations towards the UE.

UE mobility, i.e., the handover between radio cells from two different eNodeBs, is supported by the X2 interface, which is used for exchange of control signalling, e.g., handover requests, load information, and user data, i.e., forwarding during the handover process.

The interconnection to the EPC is realised by the S1 interface. Separate paths for the control information and the user data are provided by means of the S1-MME interface to the Mobility Management Entity (MME) and by means of the S1-U interface to the Serving Gateway (S-GW). As a fallback solution, handover can also be carried out by means of the S1 interface. The E-UTRAN architecture [36.300] is illustrated in Figure 2-1.

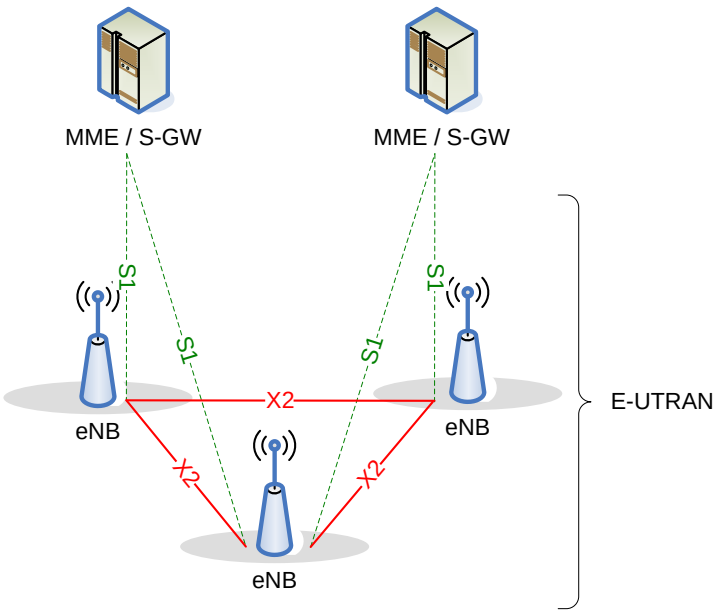


Figure 2-1: LTE overall architecture [36.300]

The functional split between the E-UTRAN and the EPC is illustrated in Figure 2-2, where yellow boxes depict the logical nodes, white boxes depict the functional entities of the control layer, and blue boxes depict the radio protocol layers [36.300].

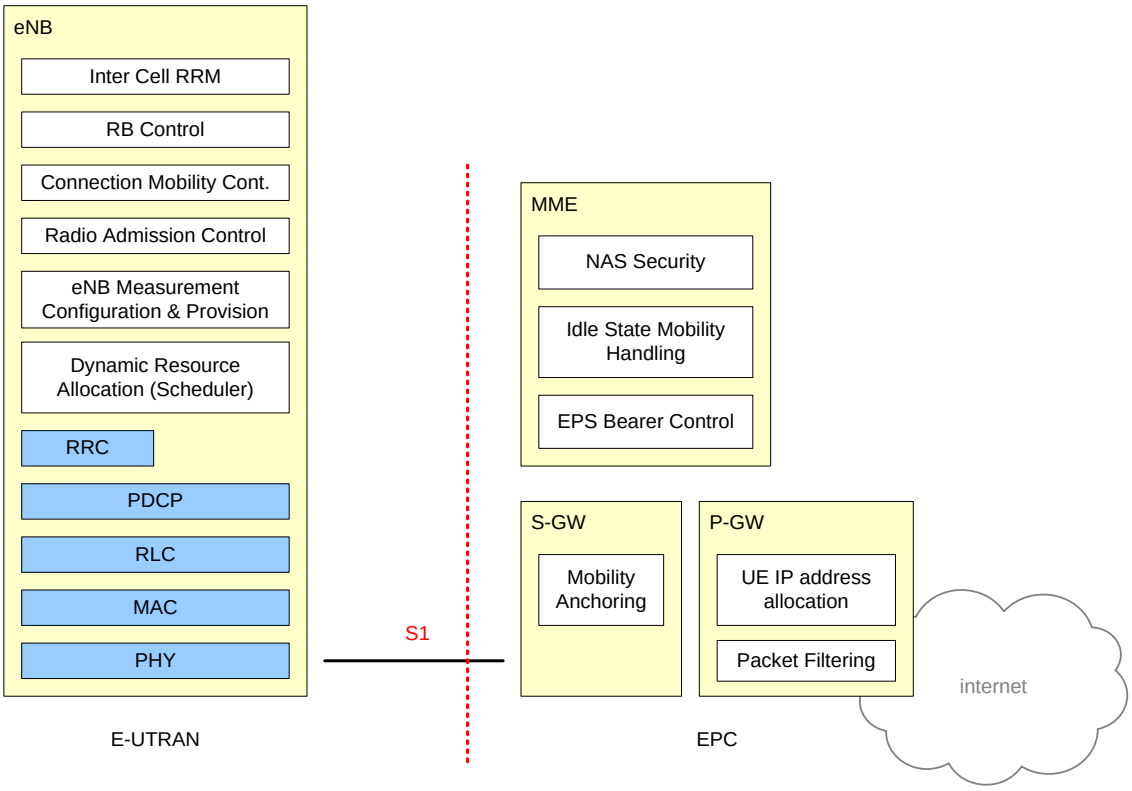


Figure 2-2: Functional split between E-UTRAN and EPC [36.300]

Starting with Release 8, a UE was only connected with one radio cell controlled by one dedicated eNodeB. Within further releases of the 3GPP specifications, i.e., LTE-Advanced, new advanced features were introduced where the UE is served by more than one radio cell:

- Coordinated multipoint transmission and reception (COMP) is targeting to ensure that the optimum performance is achieved even at cell edges. It requires very low latency between the different eNodeBs, which might be implemented within a centralised RAN (C-RAN).
- Carrier aggregation is mainly targeting to achieve higher peak data rates and to enable interference management with intelligent allocations of resources.
- With the introduction of the Heterogeneous Network (HetNet), i.e., a network composed of different base station types and different power classes, dual-connectivity was introduced.

It is envisioned that the functional architecture of 5G NORMA will provide much more effective coordination capabilities for the above mentioned LTE-Advanced features compared with the current X2 based eNodeB to eNodeB coordination.

2.3. Relevant Initiatives

Prior work on functional decomposition by other research and innovation projects is summarised in the following.

European project METIS

METIS aims to develop a 5G system concept to satisfy the requirements of the beyond-2020 connected information society, extending today's wireless communication system for new scenarios as well. In [MET-D64], METIS describes an overall, multi-facial 5G architecture which includes a numbers of Horizontal Topic concepts such as Direct Device-to-Device Communication (D2D), Massive Machine Communication (MMC), Moving Networks (MN), Ultra-Dense Networks (UDN) and Ultra-Reliable Communication (URC).

A functional architecture has been provided by identification of functionalities, which are needed by the Horizontal Topic concepts, and a functional decomposition of most important technology components proposed by METIS technical work packages. METIS provides also a logical orchestration and control architecture, which supports a flexible, scalable and service-oriented setup of NFs. Moreover, a topological and deployment architecture is provided, showing the deployments aspects and function placement options.

The functionalities derived by analysing the Horizontal Topic concept were listed and grouped into building blocks (so-called top-down analysis). A building block consists of functional elements and internal as well as external interfaces. A high-level building block depiction is shown in Figure 2-3.

Reliable Service Composition is a RAN central entity with interfaces to all other building blocks; it allows availability check and delivery of ultra-reliable links. *Air Interface* includes building blocks that are related to air interface functionalities of devices and radio nodes. *Central Management Entities* incorporates all-encompassing network functionalities, which are not exclusive to Horizontal Topics. *Radio Node Management* comprises radio functionalities, which are not specific to Horizontal Topics and that regard at least two radio nodes. Clearly, each high level building block includes sub blocks, e.g., Radio Node Management includes Mobility Management, Interference Identification and Prediction, and RAT Selection, to name only few of them). More details can be found in [MET-D64, Annex 7 and 8].

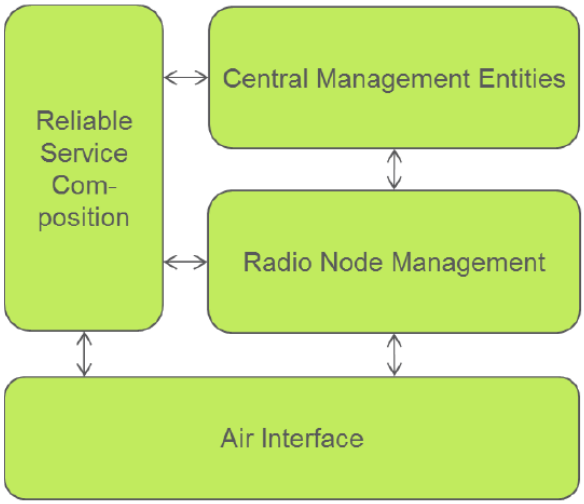


Figure 2-3: High level building blocks of the METIS system [MET-D64]

The most promising technology components provided by the METIS were analysed from an architecture perspective and thus decomposed into the so-called Functional Elements (FEs). One or more FEs may be related to dedicated Network Functions (NFs). One objective of this bottom-up analysis is to identify the constraints of NFs and to investigate the level of flexibility of functions deployment. The major result of this analysis is the *Generic Functional Architecture* where all the Functional Elements are gathered and placed in the same figure.

METIS does not address all FEs that are required to run a fully operational 5G system, such as security, authentication, authorisation and accounting (AAA), and network management and operation (OAM).

EU Project iJoin

The iJOIN project [IJOIN-D32], [IJOIN-D52] was one of the first projects to focus on flexible function allocation through the concept of RAN as a Service (RANaaS). However, in iJOIN, the function allocation was limited to deciding how to split the functionality between the edge cloud

and the central cloud (the so-called ‘functional split’). Moreover, the functional split in iJOIN was driven by the topological deployment (more specifically, the capacity and latency of the backhaul network) as well as computational demands at the central processor.

The work of the iJOIN project resulted in a few recommended functional split configurations that were similar to the ones proposed in parallel by the Small Cell Forum [SCF159]. Hence, iJOIN focused on a less flexible framework with two possible locations for NFs and without taking into account the specific service requirements when deciding the function placement.

Small Cell Forum

Small Cell Forum performed an independent study on virtualisation of NFs with reference to a LTE Small Cell framework in [SCF-15a]. A small cell is split into two components: a central small cell where functions are virtualised and a remote small cell where non-virtualised function are run. The small cell layers and functions are explored with a top-down methodology. The functions are moved from the remote small cell to the central one thus the small cell was split.

The approach is to move NFs gradually, analysing the resulting split points from the fronthaul perspective (transport latency and bandwidth requirements).

3. Flexible Network Design

The objective of the 5G NORMA mobile network architecture is to allow for integrating different technologies and enabling different use cases. Due to the partly conflicting requirements, it is necessary to use the right functionality at the right place and time within the network. In order to provide this flexibility, the network functions virtualisation (NFV) paradigm is adopted in the mobile access and core network domain, enabling mobile network functionality to be decomposed into smaller function blocks, which are flexibly instantiated.

Traditionally, mobile networks implicitly group functions into **network entities** via specification of their interconnections. Each entity is responsible for a pre-defined set of functions. Accordingly, the degrees of freedom for assigning network functionality to physical network entities are very limited. For instance, EPC elements such as gateways may be collocated with base stations in 3GPP EPS but moving only parts of the functionality of a gateway or MME to a physical base station would require proprietary modification of 3GPP interfaces. Similarly, a full centralisation of RAN functionality using the common public radio interface (CPRI) and central baseband units (BBUs) is possible. However, as such deployments use non-virtualised BBUs at the central location, it is rather relocating functionality, which does not exploit all characteristics of cloud-computing and is unable to realise e.g. pooling gains.

Replacing the traditional network of entities by a flexible "**network of functions**" allows for adapting the RAN to diverse services in a tailor-made way, by using different software rather than using just different parameterisations of the same function block. Each block may be replaceable and individually instantiated for each logical network running on the same infrastructure. Depending on the use case, requirements, and the physical properties of the existing deployment, mobile network functionality is executed at different entities within the network. Coexistence of different use cases and services would imply the need for using different VNF allocations within the network. This is enabled through the network slicing model as explained in Section 3.3. The challenge of avoiding many additional interfaces may be addressed by a flexible container protocol on user and control layer. The main benefit of the flexible functional architecture is the possibility to exploit centralisation gains where possible, to optimise the network operation to the actual network topology and its structural properties, and to use algorithms optimised for particular services, i.e., optimise through dedicated implementations instead of parameters.

The mobile network must further integrate also legacy technologies to guarantee that it can operate with existing networks. This is reflected in Section 3.1, which uses 3GPP EPS as the basis for the set of function blocks. It further adds blocks and amends existing blocks' functionalities to support 5G NORMA innovations. The novelties of the 5G NORMA innovations over the state of the art are highlighted in Section 1.3.

3.1. Functional Architecture

In this section, we provide an overview of the considered control and data layer functional architecture. The functional architecture is an essential element for meeting the 5G NORMA objectives as well as to support and integrate the novel functionalities described in Section 1.3. A detailed overview of the NFs involved is provided in Annex A. Capitalizing on that comprehensive overview, in this section, we present a concise overview of each NF, sufficient to understand their characteristics in terms of network requirements and therefore categorization and placement, as well as in terms of their interdependency among each other.

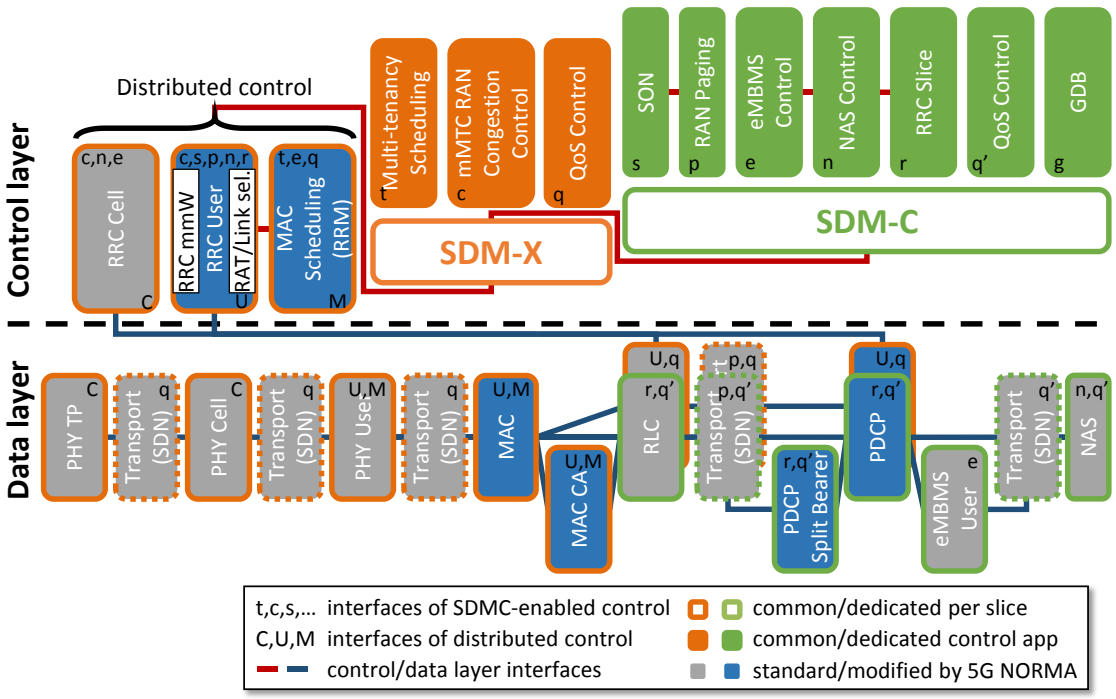


Figure 3-1: 5G NORMA control and data layer functional architecture

The c/d-layer architecture is depicted in Figure 3-1, here for the case of RAN slicing Option 2, i.e., with a common MAC layer (see Section 3.3 for more details). Functions are classified whether they belong to the control or data layer. The control layer functions are further classified into i) distributed, ii) common and iii) dedicated control. Distributed control functions are implemented as VNFs throughout the network, while common and dedicated control functions employ the SDMC concept and run as applications on top of SDM-X and SDM-C, respectively. In particular, the depicted function blocks have the following purpose:

Data layer

- **PHY Transmission Point.** The analogue and mixed signal processing for all signals transmitted (received) via one transmission (reception) point.
- **PHY Cell.** (De-)multiplexing of *PHY User* and baseband signal generation including common PHY signals for one RAT or slice.
- **PHY User.** The generation of the baseband signal (in frequency domain for OFDM-based systems) from user data (DL) and decoding of baseband signals into user data (UL), respectively.
- **MAC.** Provides functionalities such as HARQ, adaptive modulation and coding (AMC) (allow to adapt the modulation and coding to the channel quality), and discontinuous reception (DRX) (allow to improve UE battery life and energy saving).
- **MAC Carrier Aggregation.** Coordinates the exchange of scheduling information as well as feedback information corresponding to the aggregated legs.
- **RLC.** RLC Transparent Mode (RLC TM) is dedicated to forwarding RRC information to/from lower layers (RRC messages are passed unmodified to the MAC layer). RLC Acknowledged/Unacknowledged Mode (RLC AM/UM) primarily provides (re)concatenation and (re)segmentation of PDCP PDUs to adapt user data size to the amount of resources available to a radio bearer in a TTI. RLC AM provides a re-transmission process, i.e., an automatic repeat request (ARQ) service, for increased reliability.
- **PDCP Split Bearer.** Executes the functionalities of routing, reordering, and reordering timer
- **PDCP.** Carries out data transfer functions including functions sequence number maintenance, RoHC, (de-)ciphering, integrity protection, and verification.

- **eMBMBs User.** Performs the transmission of MBMS application data using the IP multicast address with the addition of SYNC protocol to guarantee that radio interface transmissions stay synchronised. Multimedia Broadcast Multicast Services (MBMS) offer support for broadcast and multicast services enabling the transmission of multimedia content (text, pictures, audio and video) and utilizing the available bandwidth intelligently [23.246].
- **NAS User.** Performs routing functionality (S-GW, P-GW) and service delivery.
- **Transport (SDN).** Connectionless routing within RAN based on SDN configuration. Within an UCA, a connectionless routing of DL small data packets from the anchor node to the best serving node (and vice versa in UL) has to be established for each UE.

Distributed control

- **RRC Cell.** Handles control layer signalling protocols associated with broadcasting system information, including NAS common information and information relevant to UEs in RRC_IDLE, e.g., cell (re-)selection parameters, neighbouring cell information, and information (also) applicable for UEs in RRC_CONNECTED, e.g., common channel configuration information.
- **RRC User.** Handles the UE management and control including radio bearer setup. In includes the subblocks
    - **RRC mmW.** It adds functionality related to mm-wave transmission points controlled by 5G coverage cell and UCA for short data packet transmission.
    - **RAT/Link Selection.** Enables link selection and packet scheduling if the UE is simultaneously connected to two or more RATs.
- **MAC Scheduling (RRM).** Scheduling the transfer of user data and control signalling in DL and UL subframes over the air interface.

Common control

- **Multi-tenancy Scheduling.** Coordinating resource sharing among multiple tenants.
- **mMTC RAN Congestion Control.** Grouping mMTC devices into context-based clusters and schedule their RAN procedures in sub-frames, to reduce the RAN congestion rate.
- **QoS Control.** Network monitoring and configuration in real time through open interfaces that interact with the SDM-X, SDM-C, and the control radio stack.

Dedicated control

- **SON.** It covers *i*) Self-configuration, *ii*) Self-Optimisation, *iii*) Self-Healing, and *iv)* User Centric Connection Area (UCA) and mm-wave cluster configuration.
- **RAN Paging.** To reduce signalling messages on the air interface and towards the CN, 5G NORMA employs a RAN paging approach, i.e. inside an UCA, in addition to a paging in a larger tracking area.
- **eMBMS Control.** Performs the admission control and allocation of the radio resources, UE counting procedure, MBMS session management (initiating the MBMS session start and stop procedures), allocation of an identity and the specification of QoS parameters associated with each MBMS session.
- **NAS Control.** It includes the subblocks
    - **NAS UE specific.** Refers to the user specific NFs and procedures related to the data layer, which are triggered by the NAS UE-specific control layer functions.
    - **NAS UE Specific and Data Layer.** Refers to the user specific functions and procedures related to the d-layer, which are triggered by the NAS UE-specific c-layer functions.
    - **NAS UE Specific and Control Layer.** Refers to the user specific functions and procedures related to the non-radio signalling between the UE and MME.
    - **NAS Event-Control Layer.** Refers to the network-side c-layer functions and procedures including those of the interface between RAN and CN (S1-C in 3GPP LTE [36.300]), which are provided to facilitate mobility management.

- **NAS Core Network.** Refers to network management functions, which are provided to support O&M functions of 5G CN.

- **RRC Slice.** Handles the UE management and control related to the slice specific part of the RAN protocol stack.
- **GDB.** Geolocation Database (GDB) stores information linked to geolocation, and makes decisions based on that geolocation information.

Beyond the classification into data layer and one of three control layer flavours as detailed above, NFs have been characterised regarding additional criteria, namely, whether synchronous or asynchronous operation with respect to TTI takes place; whether they are applied on a cell, user, or bearer granularity level; whether and which reports are generated; whether they are agnostic to the employed RAT; and their dependency to other NFs. This characterization is intended to be useful for both current LTE and future 5G New Radio (NR). Table 3-1 shows these characteristics for all of above function blocks.

Table 3-1: Properties of considered function blocks

Function block	Grouped functions	Service dependent	TTI Sync	Granularity	Reports	RAT anostic
Data layer						
PHY TP	A/D; Signal generation; (De)modulation; MIMO precoding/equal	No	Y	TP	Y/N	Y
PHY Cell	Resource mapping; MIMO mapping	No	Y	Cell	Y/N	N
PHY User	FEC	No	Y/N	UE	N	N
MAC	HARQ, AMC, DRX, UE Power control; Padding; Multiplexing of TBs; UE radio network ids	No	Y/N	UE	Y/N	Y/N
MAC CA	Scheduling info exchange; common priority handling; UL coordination	No	Y	UE/ Cell	Y/N	N
RLC	TM: Buffering/transferring of PDCP PDUs; AM/UM: Concatenation/Segmentation; Reordering; Duplicate detection PDU; Reassembly SDU; AM: Error correction ARQ; Re-segmentation	No	N	Bearer	N	N
PDCP Split Bearer	Reordering timer; Routing; Reordering;	No	N	Split Bearer	Y	N
PDCP	Data transfer; Seq number maintenance; ordered delivery and duplicate detection; discard timer; Integrity protection; (De)ciphering; SL-(De)ciphering; ROHC	Yes	N	Bearer	Y/N	N
eMBMS User	MBMS application data distribution to TPs, SYNC Protocol	No	N	MBSFN Area	Y	N
Transport (SDN)	Paging controlled by anchor eNodeB of the UCA. Less cells compared to MME paging	UE with small data packets (UCA)	N	Area of cells	Y	N
Distributed control						
RRC Cell	Broadcast System Information	SI	N	Cell	Y/N	Y
RRC User	RRC connection mgmt.; Paging; RB mgmt.; Connection mobility; QoS management functions; Security, Measurement configuration and reporting	Connection control and Reporting	N	User	Y/N	Y
RRC mmW (RRC User)	mm-wave support	No	N	User	Y/N	Y
RAT/Link Selection (RRC User)	Inter RAT link Selection, Inter RAT packet scheduling, Multi-connectivity to UE.	Yes	N	User	Y	Y
MAC Scheduling (RRM)	(e)ICIC, CoMP; intra-site; inter-site; intra-cell; priority handling; channel mapping; scheduling; BS power control; D2D support	No	Y	Cell	Y	N
Common control						
Multi-tenancy scheduling	Network resources sharing to different slices/tenants.	No	N	Cell	Y	N

mMTC RAN cong. control	Handling massive machine type access	No	N	Cell	Y	N
QoS Control	QoS control function on common control	Y	N	Bearer/Service	Y/N	Y
Dedicated control						
SON	CCO, Energy, MRO, PCI configuration, 5G cluster configuration	UCA and mm-wave transmission	N	Area of cells	Y	N
RAN Paging	Paging controlled by anchor eNodeB of the UCA. Less cells compared to MME paging	UE with small data packets (UCA)	N	Area of cells	Y	N
eMBMS Control	Admission control and allocation of the radio resources, Counting, MBMS Session management	No	N	MBSFN Area	Y	N
NAS UE-U (NAS Control)	Mobility anchor mgmt.; packet routing; packet marking; deep packet inspection; DHCP; data forwarding in RAN; lawful interception; UL/DL charging	No	N	UE or bearer	Y/N	Y
NAS UE-C (NAS Control)	Inter-node signalling for mobility; NAS signalling to UE; NAS security; NW attachment; Authentication; TA mgmt.; Charging; Paging; RAN info mgmt.; context transfer HO; GTP mgmt.; Bearer mgmt.; packet forwarding	No	N	UE or bearer	Y/N	Y
NAS Event-C (NAS Control)	PGW/SGW selection; MME selection for handover; UE reachability procedure; location reporting; S1 UE context mgmt.; ERAB mgmt.;	No	N	Event/UE	Y/N	Y
NAS Core Net (NAS Control)	GTP-C load control; MME load balancing; MME overload control; PDN GW overload control	No	N	CN/UE	Y/N	Y
GDB	Geolocation-database based heterogeneous connectivity; resource management support	No	N	Cell/UE	Y	N
QoS Control	QoS control function on dedicated control (similar to counterpart on common control)	Y	N	Bearer/Service	Y/N	Y

Table 3-2 lists for each of these blocks the 5G NORMA innovations from D4.1 Part II that impact the corresponding function block.

Table 3-2: Impact of partner contributions to function blocks

5G NORMA Innovation	User-Centric Connection Area	RAN supports for providing optimised on-demand adaptive network functions …	Mobile Edge Computing and NW RA for Multi-Tenancy	Multi-Service Technologies	Multi-tenant dynamic resource allocation	Multi-RAT integration	Data-layer and control-layer design for multi-connectivity	RAN support for flexible 5G service-flow (SF) concept with in-SF QoS differentiation …	Multiple connectivity at the different layers	Centralised Radio Resource Management	Virtualised Geolocation Databases for Heterogeneous Network Management and Optimisation	Functional (de)composition for Supporting Multi-Tenancy in HetNets	Multi-Tenancy in Multi-RAT Environments	Load Balancing of Signalling Traffic	QoS innovation
Partner→ ↓Function block	**ALU**	**Nokia**	**NEC**	**TUKL**	**UC3M**	**ALU**	**Nokia**	**Nokia**	**TUKL**	**Nomor**	**KCL**	**NEC**	**UC3M**	**KCL**	**Atos**
Data layer															
PHY TP						•									
PHY Cell		•	•			•									
PHY User		•	•									•		•	
MAC	•	•	•								•			•	
MAC CA								•			•				

RLC		•						•				•			
PDCP Split Bearer		•				•	•	•		•					
PDCP	•	•		•		•		•	•		•				
eMBMS User															
NAS															
Transport (SDN)	•										•				
Distributed control															
RRC Cell		•								•	•				
RRC User	•	•	•			•		•	•	•	•	•	•	•	•
RRC mmW	•					•					•				
RAT/Link select.									•		•				
MAC Scheduling	•		•	•	•	•			•	•	•	•	•	•	•
Common control															
Multi-tenancy Sch.					•								•		
mMTC Cong. Ctrl				•											
QoS Control															•
Dedicated control															
SON	•					•					•	•	•		
RAN Paging	•														
eMBMS Control															
NAS Control	•		•	•											
RRC Slice								•							
QoS Control															•
GDB											•				

3.2. SDMC Interfaces

The 5G NORMA functional control and data layer incorporates the novel concept of software-defined mobile network control (SDMC). In the following, the interfaces of those novel centralized SDMC-enabled control functions, running as applications on top of the SDM coordinator (SDM-X) or SDM controller (SDM-C), will be considered in more detail. The interfaces enable the SDMC apps to control the "legacy" distributed control functions as well as the distributed data layer functions.

The SDM-C and SDM-X configure the 5G network architecture including NFs and SDN transport elements via their southbound interface (SBI). An SBI provides an abstraction of the NF to the SDM-C/X, enabling direct representation of the NF behaviour and requirements. The SDMC applications presented below shall require a specific set of information to operate. The SDM-C/X extract such information from the distributed data and control layer NF via the SBI. The list of SBI requirements is summarized in Table 3-3.

Table 3-3 SBI requirements from SDMC applications

SBI requirements	SDMC application
Abstraction of underlying topology and NFs	Required by all applications
Multi-tenant and multi-slice monitoring of QoS and resource usage	Multi-tenant scheduling, QoS Control
Dynamic configuration of the data layer transport functionality (SDN)	RAN paging, NAS Control
Monitoring of c-layer specific information	Required by all applications
Configuration of network functions related parameters like scheduling policies, QoS values, etc.	Required by all applications

Support access to applications with low latency requirements hosted close to the access network within the operator trust domain	NAS Control

3.2.1. mMTC RAN Congestion Control

The mMTC RAN congestion control SDM-X app groups mMTC devices into context-based clusters and schedules their RAN procedures in sub-frames to reduce the RAN congestion rate. For this purpose, it connects to RRC Cell RRC User.

Information exchanged over "c" interface:

C-layer information related to mMTC device clustering is exchanged between mMTC RAN Congestion Control (mMCC) and RRC Cell/User over the "c" interface. The information includes:

a) UE context information, such as device class and geo-location, to select a device group for every UE,
b) grouping result, including the device role in its group and the group controller identity to support D2D intra-group communication, and
c) requests, confirmations, and failure reports

Requirements in terms of latency and bandwidth:

The execution frequency of mMTC RAN congestion control processes depends on the level of UE mobility and dynamics, but is generally low. Hence, a latency up to a few seconds is acceptable. The bandwidth requirement depends on the specification of group size. The C-layer signalling overhead size grows almost linearly with the group size, with each additional group member several extra bytes should be considered in the broadcasted grouping result.

Scalability:

The mMTC RAN congestion control concept is limited in scalability as it focuses on the RAN congestion in local cells and highly relies on the D2D connections between UEs.

3.2.2. SON

The SON functionality in LTE defines optimized measurement parameters such as time to trigger or handover margin to reduce handover failures, i.e., radio link failures during mobility of a UE. These parameters are provided by RRC signalling towards each UE. The setup and update of a Neighbour Relation Table (NRT) is also part of SON. For each eNB, such a NRT is defined to check possible Physical Cell ID (PCI) confusions and collisions. Within 5G NORMA, a clustering of access nodes is proposed to reduce signalling towards the CN (UCA concept) or to improve the mm-wave transmission quality with respect to blocking effects caused by sudden user movement or obstacles entering the transmission path.

Information exchanged over "s" interface:

C-layer information related to setup of a UCA cluster or a mm-wave node cluster is transferred from the RRC mmW control to the SON control. This information includes

a) UE measurement of neighbour cells belonging to other access nodes to setup a UCA or mmWave cluster,
b) the indication that a UE has selected a better serving access node within the UCA cluster,
c) the indication that for a UE a new UCA has to be defined in case the UE leaves the defined UCA cluster. This indication includes a new best serving access node outside the UCA cluster and UE measurements of other access nodes.

The SON control will define a UCA or a mm-wave cluster based on these inputs and other internal information such as the NRT or even anticipatory techniques such as mobility tracks of UEs. Each cluster is always specific for one individual UE.

Requirements in terms of latency and bandwidth:

This SON functionality requires low computational load as only limited amount of data is provided by the RRC mmW control. This functionality is asynchronous as it is either triggered once if a UE is addressing the 5G system or by UE mobility inside the UCA or the mm-wave cluster.

In case of mobility, i.e., a new UCA or mm-wave cluster has to be defined, the latency of these SON processes should be in the order of an envisaged 5G handover process, as the definition of a new cluster should be included within the RRC handover message towards the UE. In case the latency, e.g., based on processing delays, is too high, RRC User can still define a new UCA or mmWave cluster based only on the cells of the new best serving node and can update the cluster later on with an additional RRC message including the cluster defined by the SON control.

Scalability:

For both the UCA cluster and the mmWave cluster, the definition inside the SON process is UE specific, based on the UE measurements and e.g. anticipatory information. Therefore, multiple instances have to be set up, processed, and maintained. As a cluster has to be defined during initial access of an UE and during mobility, the processing capabilities of these SON processes should take into account the envisaged access rate and mobility parameters of UEs.

3.2.3. RAN Paging

The RAN paging functionality is responsible to perform a paging request inside the UCA cluster, i.e. a paging within RAN without involvement of the CN. The selection of the access nodes inside the UCA is done by the SON process, which provides for each UE the UCA information, i.e., the anchor node, the best serving node, and other nodes comprising the UCA.

A UE moving inside the UCA may indicate a new best serving cell by using a 1-step protocol [SWS15], which minimizes the air interface load but may be lost as no RLC with retransmission is involved. In case the UL notification of a new best server was not received, the last serving node will not receive an UL acknowledgment notification related to the DL transmission. In this case, the last serving node will inform the anchor node, which will request the start of a UCA paging via the RRC mmWave towards the RAN paging functionality. Based on the UCA information from the SON process, i.e., an internal interface between SON and RAN paging, the involved access nodes are provided by the SON process.

Another task of the RAN paging is the setup of the connectionless transport in UL and DL direction inside the UCA by configuring the Transport (SDN) functionality.

Information exchanged over “p” interface:

C-layer information related to a paging request in the UCA cluster is transferred from the RRC mmWave to the RAN paging control. This information includes e.g. the UE_ID or the UCA_Id to identify the UCA cluster. The access nodes belonging to the UCA are provided by the SON process.

C-layer information related to paging from the RAN paging control to RRC mmW. This information includes the access nodes, which have to initiate paging, and the related the UE_ID.

C-layer information related (access nodes) to setup of connectionless data paths for DL and UL within the UCA cluster towards the Transport (SDN) functionality.

Requirements in terms of latency and bandwidth:

This RAN paging functionality requires low computational load as only limited amount of data is exchanged between RRC mmW and RAN paging and between RAN paging and Transport (SDN). This functionality is asynchronous as it is triggered if the serving node inside the UCA is not known or a UE selected a new best serving node.

Latency with respect to RAN paging is not critical, the setup of the connectionless data paths should be in the order of an envisaged 5G handover process.

Scalability:

Both, the UCA paging and the setup of the connectionless data paths are UE specific. The process has a limited life, i.e., as soon as paging or the setup is acknowledged, the process can be terminated.

3.2.4. eMBMS Control

Identifying quality issues of eMBMS in a certain area is a big challenge for an operator. In fact, for eMBMS, no feedback information from the UEs is available but it is important for the operator to make sure that the users receive the service in an acceptable quality.

Service providers need configure the most optimal parameters for a certain MBSFN area. The coverage requirement, for example, needs a proper selection of the modulation and coding scheme (MCS) for that area and spectral efficiency is directly related to the MCS selected for the transmission. Choosing a low MCS for eMBMS data will accomplish coverage at the cost of precious unicast resources.

In [ABK+10], different approaches for the MCS selection have been proposed for MBSFN transmission over a LTE network. The approaches cover different needs such as the assurance of service continuity for the user with lowest SINR value, the selection of the MCS that maximizes the spectral efficiency, the selection of MCS based on the covered area, or the fraction of users that receive the service in an adequate quality.

Hence, one example for a possible task of the SDMC-enabled function block, respectively SDMC application *eMBMS Control* specified in section 6.20, could be the configuration of different MCS selection policies for an MBSFN transmission.

Information exchanged over "e" interface

The eMBMS Control function block can be considered as an application, which is responsible for the configuration of the c-layer NFs involved in eMBMS service distribution (RRC Cell, MAC Scheduling, eMBMS User, NAS Control). With reference to the example mentioned above, the the SDMC app receives through SDM-C the target KPI requirement for the MBSFN transmission provided by the service provider. The SDMC app selects the most efficient MCS selection policy and the proper configuration is signalled through the SBI to the MAC Scheduling NF.

Requirements in terms of latency and bandwidth

eMBMS Control is asynchronous with respect to the TTI and handles only c-layer, hence, it does not impose strict timing requirements.

Scalability

There is one eMBMS Control block per MBSFN (Multimedia Broadcast Single Frequency network), talking to every instance of the involved c-layer NFs.

3.2.5. NAS Control

The subblocks of *NAS Control* are described in Section 6.21. In LTE, NAS functions are provided by the MME for control layer and S-GW/P-GW for data layer in the NAS. Using SDM-C and SDM-X, 5G NORMA allows for flexible and scalable on-demand NAS functions such as mobility management and QoS control with possibly distributed c/d-layer gateways. MME, QoS Control, and SON in 5G NORMA, as shown in Figure 3-1, are considered as applications of SDM-C or SDM-X. NAS Control may be implemented as part of SDMC applications and interacting with, e.g., RRC Cell, RRC User, and NAS data layer over south-bound interfaces of SDM-C or SDM-X, referred to as "n" interface in Figure 3-1.

Information exchanged over "n" interface:

C-layer information related to, for example, connection and mobility management, session management for individual idle and connected UEs, as well as core network management, network operation, and lawful interception.

Requirements in terms of latency and bandwidth:

The following general requirements of 5G networks [23.779] may have direct implications on the "n" interface:

- Support services that have different latency requirements between the UE and the Data Network.
- Minimize the signalling (and delay) required to start the traffic exchange between the UE and the data network, i.e., signalling overhead and latency at transition from a period where UE has no data traffic to a period with data traffic.
- Support access to applications (including 3rd party applications) with low latency requirements hosted close to the access network within the operator trust domain.
- Support optimised mechanisms to control signalling congestion (includes avoidance).
- Efficient network support for a large number of UEs in periods without data traffic.

Scalability:

In legacy LTE networks, the network elements, which provide NAS Control including MME, S-GW, and P-GW have to manage scalability independently. The SDM-C and SDM-X based NAS Control over the "n" interface allows for utilizing distributed computation using cloud based scaling techniques or enablers such as:

- The subscription of UE contexts as well as reachability control data of individual users can be maintained largely independently, enabling massively parallel processing and handling of UE contexts;
- NAS Control can be provided by a highly scalable control-layer functional entity of which each instance is capable of providing and managing all the functions needed for handling a single UE's state management, as opposed to having at least MME, S-GW and P-GW involved;
- The same subscriber context database is accessible and modifiable for all UE context related control layer functions (MME, SGW, PGW, PCRF) or a centrally available context database can be used;
- The virtual state machine of the control-layer functional entity, which is handling NAS Control for a set of UEs, need to be aware of UE states only during active transaction. Hence, there is a trade-off between the statelessness and database access frequency (user state is removed in between the transactions).

3.2.6. GDB

Location is fundamental to many aspects of mobile provisioning. It is vital to the configuration and parameterisation of the mobile network and devices. It is further essential to the provisioning of computational and other (e.g., "bare metal") resources in order to realise a softwarised mobile network in a virtualisation context. The GDB is therefore fundamentally related to and interfaced with the SDM-O and SDM-X, as well as the instantiations of mobile network elements.

In the context of mobile network and device functional capabilities, the GDB has numerous purposes. These can both augment capabilities that are already present in mobile networks, e.g., awareness of signal powers and related assumptions in an RRC context might be augmented by location information, and present entirely new capabilities that otherwise would not be viable, e.g., satisfying requirements linked to propagation path in latency-sensitive applications.

Information exchange over "g" interface

Information exchange is extremely varied given the applications to which the GDB applies. Nevertheless, the following are the key information sets envisaged.

- Location information, as obtained using satellite-driven positioning systems, potentially also augmented (with reliable confidence bounds) by terrestrial solutions such as triangulation, TDOA/AOA, broader A-GPS, and other methods. This information is relevant and can be provided at many different levels, depending on usage purpose of the GDB: RAN-level, UE-level, at the level of radio elements supporting mmW
- Resource availability information linked to the location information, again being relevant at many different levels mirroring the location information above. There are many examples here, e.g., radio resources including fine-granular subsets, and computational and bare metal resources.
- Device (e.g., base station, UE, RAN, etc.) identities needed for internal network operational purposes and techno-regulatory solutions
- Information on chosen resources, again, particularly relevant in the context of advanced techno-regulatory solutions, but also needed for internal network operations.
- Device/network capabilities and characteristics (including bare metal)

Requirements in terms of latency and bandwidth:

The actual requirements vary widely and depend on the chosen purpose of the GDB, noting again that the GDB can serve many purposes. Typically, very high latency and low bandwidth provisioning would be sufficient, and this is the case also for the context where novel techno-regulatory related mechanisms might be realised, e.g., to access more spectrum, such as akin to LSA, CBRS, TV white spaces. Such cases could cope with a latency of a second or more, and a data load for each communication with the GDB of typically less than 1 kB or a small number of kB. Such latency can also be sufficient for some of the internal mobile-network applications of the GDB, such as RRC on a large-scale (e.g., for slices, upon the formulation of slices), use of the GDB for propagation path reduction in creation of edge-slices for ultra-low latency applications (such as the Tactile Internet). However, that assumptions on low bandwidth depend on the number of devices or elements accessing the GDB (see the discussion under "Scalability").

For other purposes that the GDB might serve, such as assisting the MAC, RRC mmW and others, it is likely that a high bandwidth and low latency would be required.

Scalability

The GDB concept is highly scalable, although the extent to which it must be implemented as such is dependent on the intended application. The GDB can operate from at the very small scale (e.g., within devices themselves—perhaps in a distributed fashion) up to, at a large scale, across the mobile network or even across many different systems looking to access common spectrum (e.g., at the regulatory or regulatory-certified level).

The use of cloud-based and distributed computing techniques in order to handle scalability challenges, particularly in the context of techno-regulatory related schemes, for example, is entirely viable.

3.2.7. QoS Control

The QoS control function block can be seen as an application related to the control of the services implemented on 5G NORMA. It uses the functionalities and APIs offered by the SDM-C/X northbound interface (NBI), and the radio control block (MAC scheduling) in the control layer, to manage the QoS in the network.

The QoS control function block is an application composed of two basic algorithms that enable the QoS monitoring and the QoS enforcement. The first algorithm is in charge of the QoS parameter set that has to be monitored and also receiving the events captured when some value is not correct. The second algorithm is oriented towards the control and evaluation of the QoS values during the service lifetime and to enforce actions if needed.

Information exchange over "q"/"q'" interface

The QoS control function block will interact with the SDM-C/X and the radio control functions through the q/q' interface. The SDM-C/X NBI has to provide methods to the QoS control function block to carry out the following actions.

- QoS monitoring. Control of the QoS parameter set that has to be monitored through the SDM-C/X and radio control blocks, and reception of the events reported by the NFs. An event is reported when some of the QoS parameters evaluated are not fulfilling the requirements.
- QoS enforcement. Control and evaluation of the QoS values over the NFs placed in the network, based on the monitoring information received and the constraints defined by the service provider. If new actions (parameter reconfiguration, resource re-schedule) are required, new configuration information is sent.

Requirements in terms of latency and bandwidth

The QoS control function block does not have strict bandwidth requirements, i.e., moderate latency and bandwidth is sufficient to manage the data exchanged between QoS control function block and SDM-C/X and between QoS control function block and radio control block (MAC scheduling). This function is asynchronous and is related to a specific service as well as associated events.

On the other hand, a low latency is required in the case that data and events should be transferred almost instantaneously without experiencing a noticeable delay. Very sensitive information is exchanged through this interface and the latency is crucial to fulfil the service requirements.

Scalability

The QoS control function block is service specific and depends on the QoS parameter measurements and the constraints defined per service. Therefore, multiple flows have to be monitored and multiple events could be triggered. As this application is per service, different configurations could be used with different QoS values based on the service provider requirements, meaning that the function will manage different parameters per each service configuration.

3.2.8. RRC Slice

RRC Slice function block is described in Section 6.22. RRC Slice is considered as part of dedicated control layer functions and may be implemented as an application on top of SDMC. RRC Slice is interacting with at least RRC User over SDMC south-bound interface, referred to as "r" interface in Figure 3-1.

Information exchanged over "r" interface:

Details to be included in D4.2.

Requirements in terms of latency and bandwidth:

RRC Slice shall not introduce considerable additional latency and signalling overhead on top of RRC User.

Scalability:

SDMC based RRC Slice allows for utilizing distributed computing using cloud based scaling techniques or enablers, as described for other SDMC based applications such as NAS Control.

3.2.9. Multi-tenancy Scheduling

The multi-tenant scheduling functionality is responsible to control the underlying scheduling for allocating dynamically resources to different slices. A new tenant can send a request to this module in order to obtain the amount of resources according to the service requirements. When a new request has been received, the multi-tenant scheduling application, given the actual network load information, decides if the request is rejected or accepted. In the latter case, it controls the MAC scheduling (RRM) through the SDM-X in order to serve the tenant's users properly.

Information exchanged over "t" interface:

C-layer information required to allocate the network resources to different slices. It is part of the interface with SDM-X to provide control information regarding the policies to share resources among slices and to receive information regarding network load. Through this interface, it controls the MAC Scheduling (RRM) function.

3.3. Network Slicing

One of the major elements of the 5G NORMA architecture is network slicing [5GN-D31]. In the following, we refer to a network slice as a separate logical mobile network which delivers a set of services with similar characteristics and is isolated from other network slices. Different slices may eventually share the same set of requirements, e.g. towards the radio access network and could therefore use the same radio access network functionality with possibly different configuration per service but without instantiating individual implementations per slice. This is already shown in [5GN-D31] where each slice may have dedicated resources but also share resources (common resources) with other slices.

In this section, we highlight the potential of network slicing as an enabler for flexible and scalable mobile networks. In fact, the concepts of *flexibility and scalability* are imperative for ensuring that mobile networks can be appropriately adopted to network environments of a particular use case. This allows for cost- and energy-efficient deployments which are not targeted to individual mobile network solutions but are rather versatile enough to accommodate a variety of requirements into one solution. In this regard, network slicing allows for providing customised logical mobile network instances which suit each individual application.

3.3.1. Implementation of RAN and CN Slicing

3.3.2. RAN Slicing Options

In this section, we elaborate on three options of RAN slicing which are illustrated in Figure 3-2. This highlights how the different aspects of shared RAN slices can be integrated.

1) The first option (Option 1) shows two network slices where each carries two different services. Each slice may be operated by a different mobile network operator (MNO). Furthermore, for each slice an individual RAN protocol stack is implemented down to the upper part of the physical layer (c.f. PHY-User and PHY-Cell). Only the lower part of the physical layer (c.f. PHY-TP) is shared across slices. The multiplexed access to PHY-TP is coordinated by the SDM-X which makes use of flexible and efficient radio resource management, such as in-resource and user-centric control, where different numerologies are supported within the same spectrum. One could think of Option 1 as implementing all user-specific functions such as forward error correction encoding, layer mapping and precoding in an individual fashion, while TP-specific functionality such as transmission of synchronisation and cell-specific reference signals are shared.

2) Option 2 depicts again two network slices from two operators. Compared to the previous example, each slice uses an individual implementation of service-specific functionality such as PDCP, RLC, and slice-specific RRC. In addition, the tenant may implement a

customized QoS scheduling to perform pre-scheduling. The access to the MAC layer is then controlled by the SDM-X where resource fairness across tenants and QoS guarantees corresponding to individual SLAs must be met. Furthermore, resource isolation must be provided to alleviate side-effects.

3) Option 3 illustrates the case of two operators using the same RAN as shared resource, i.e., the SDM-X is the interface between CN and RAN. In this example, no customisation of radio resource management beyond SDM-X parameters and configuration would be possible.

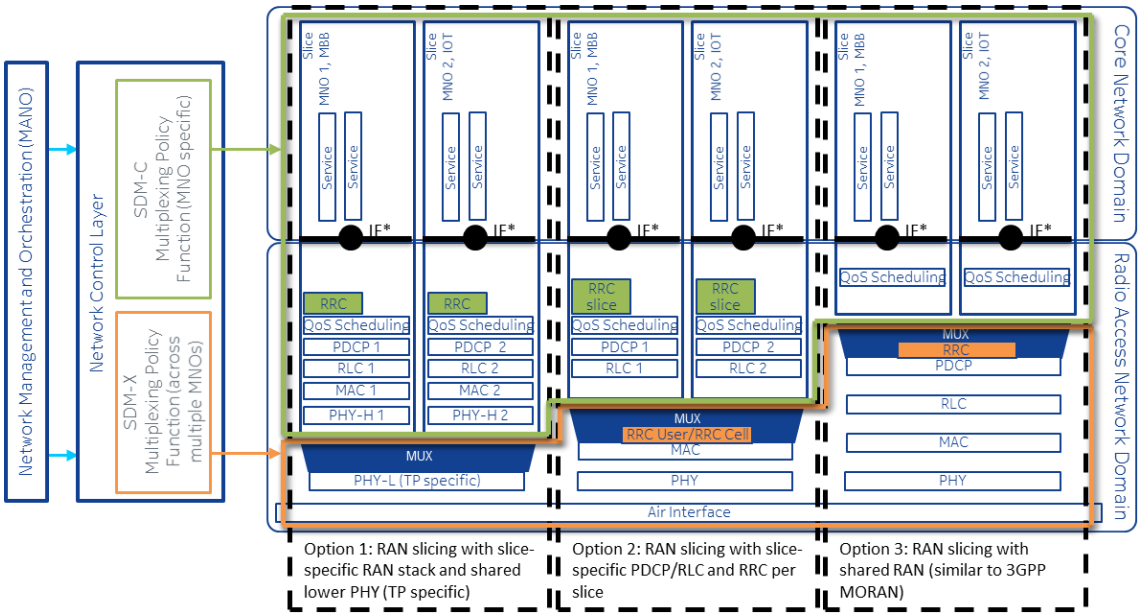


Figure 3-2: General architecture option including NW slicing and multi-connectivity

3.3.3. Realisation of network slicing in CN and RAN

Figure 3-3 illustrates how RAN slicing and CN slicing can be integrated to ensure an end-to-end network slicing (here shown without MANO and Network Control Layer as in the previous figure). As a basis, RAN Slicing Option 2 is applied. The example in Figure 3-3 differentiates the control and data layer of the core network, i.e., all data layer functions are implemented specifically for each slice. By contrast, control layer functionality may be divided into those functions which are applicable to all slices (e.g., Mobility Management) and those which may be customized for each slice (e.g., AAA). Hence, a common entity first receives all control layer signalling which is then either directly processed (common functions) or forwarded to the corresponding slices (dedicated functions).

Referring to Option 2, each data layer context then maps to an individual RAN slice that uses customized PDCP, RLC, QoS Scheduling, and slice-specific RRC, which are implemented service-specific. Then, the MUX may apply user-specific or cell-specific RRC which would map to the common CP functions (e.g. Mobility Management). Furthermore, the same communication interface may be used between the RAN, common CN control, and specific CP in order to allow for adaptability and flexibility of function location.

Referring to Option 2, each data layer context then maps to an individual RAN slice that uses customized PDCP, RLC, QoS Scheduling, and slice-specific RRC, which are implemented service-specific. Then, the MUX may apply user-specific or cell-specific RRC which would map to the common CP functions (e.g. Mobility Management). Furthermore, the same communication interface may be used between the RAN, common CN control, and specific CP in order to allow for adaptability and flexibility of function location.

The RAN is a typical example of a shared network function controlled by a single authority, where spectrum is shared amongst mobile virtual network and service operators. As shown in Figure 3-3, the c-layer is split into cell related functions which are common to all slices, and session or user specific RRC. Depending on the underlying service, RRC can configure and tailor the d-layer protocol stack. For example, for a slice supporting low delay services IP and related Header Compression (HC) may not be used, and RLC can be configured in transparent mode. In contrast, for services requiring QoE and excellent QoS, IP as well as acknowledged RLC must be initiated. In addition, there would be the possibility to chain proprietary and operator specific functions within a network slice. In this regard, the intra-slice application (QoS) scheduler (which prioritises sessions within the related slice) is customized per slice, while the inter-slice radio scheduler (which schedules different slices) resides in the common RAN part. Multi-service scheduling is part of a flexible RAN and provides the capabilities to differentiate traffic classes and assign resources according to QoS requirements. Hence, service flows from different slices can be individually treated, e.g., flexible numerologies can be used to fulfil QoS constraints and even semi-persistently reserved resources for deterministic traffic requirements.

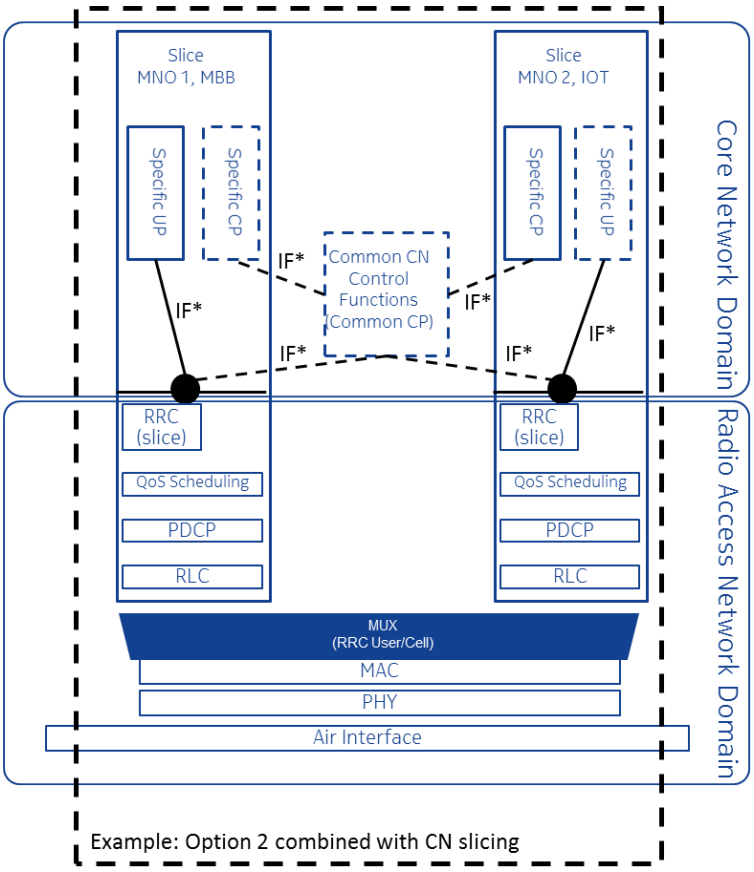


Figure 3-3: Integration of RAN slicing and CN slicing

The RAN slicing architecture according to Option 2 is shown again in Figure 3-4 using the functional architecture introduced in Section 3.1 and based on the architecture logical view introduced in [5GN-D31]. Compared to the above figure, this view separates control and data layer. On data layer, PHY-TP, PHY-Cell/User, and MAC would be common functions, while upper RAN protocol functions (RLC, PDCP) are customized per slice. Furthermore, session management may be customized per slice. On control layer, RRC Cell/User, MAC scheduling, and Mobility Management are common control layer functions, while RRC Slice and AKA may be customized. In this case, each slice may use its own AKA including authorization and encryption, which is then applied by the dedicated data layer functions. Hence, a cryptographic isolation of slices would be enabled. Furthermore, on control layer, shared functions such multi-tenancy scheduling and QoS

control may be coordinated by SDM-X, and dedicated functions such as SON and QoS scheduling may be implemented specifically for each slice.

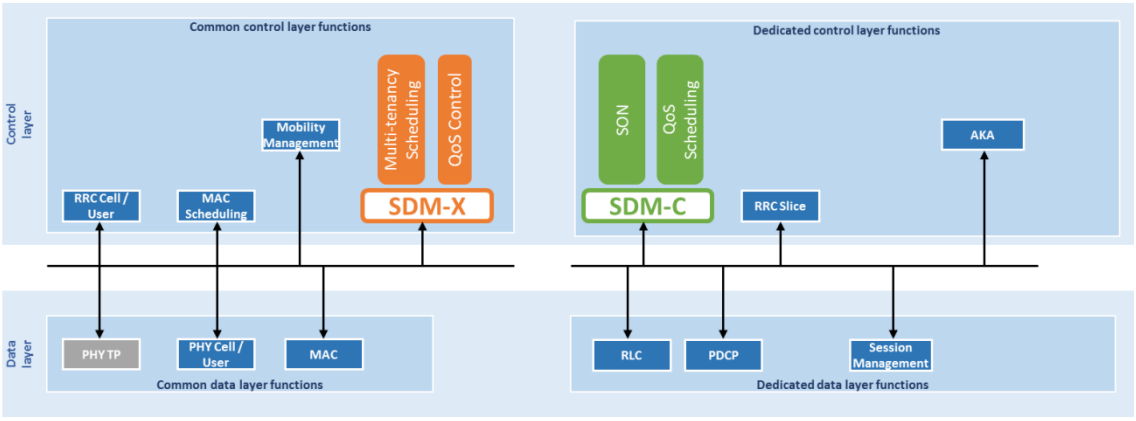


Figure 3-4: RAN slicing Option 2 using the functional architecture introduced in section 3.1 and the architecture logical view

3.3.4. Deployment options

5G NORMA considers three different deployment options, i.e., bare metal, edge cloud, and central cloud, which are detailed in [5GN-D31] using the architecture deployment view. Figure 3-5 illustrates the application of RAN slicing based on the functional architecture introduced in Section 3.1 and applied to the architecture deployment view. If we compare this with the slicing architecture in Figure 3-3, we can see that the common network functions in RAN such as PHY and MAC would be deployed either as physical NF (PHY-TP) or as VNF at the edge cloud, e.g., MAC may be implemented as PNF in the case cMTC or as VNF in the case of MBB services. In either case, the proximity to the radio access point is important as those function blocks are operated with hard real-time constraints. Higher layers, which are slice-specific, may be deployed in the edge cloud in order to keep proximity to the radio access point but also relaxing constraints on the transport network. This is possible because those functions are not subject to hard real-time constraints. Hence, the division into dedicated and common functions in RAN slicing option 2 would be reflected by the deployment architecture.

The same deployment architecture is also applicable to RAN slicing option 1 where most of the RAN protocol stack is customized. However, in that case those functions would be executed as VNFs within the edge cloud rather than as PNFs on bare metal.

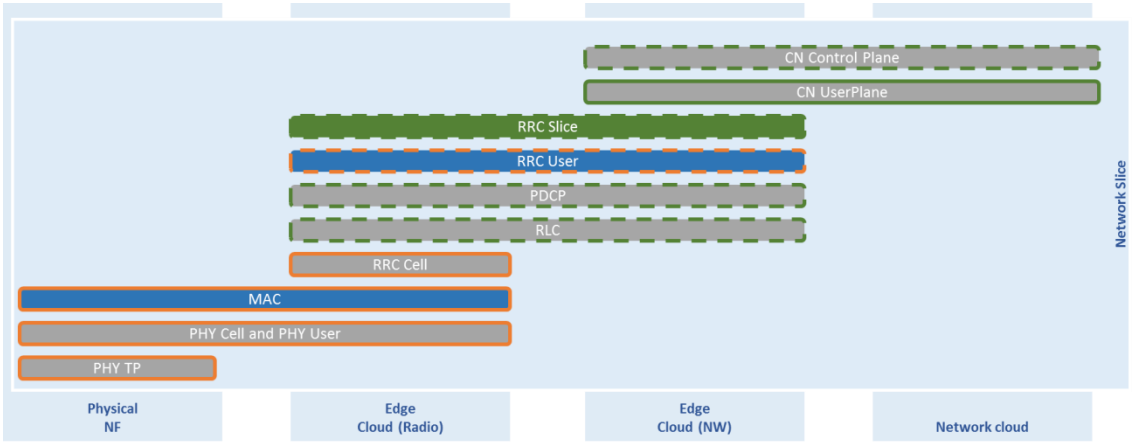


Figure 3-5: Application of RAN slicing Option 2 based on the functional architecture in section 3.1 and applied to the architecture deployment view

3.3.5. Integration of RAN Slicing and Flexible RAN Technologies

A key element of the 5G NORMA architecture is the support of heterogeneous and flexible networks with respect to radio access technologies and deployments. The control of this heterogeneous network and the orchestration of the function assignment is a central element of the SDMC concept developed by 5G NORMA. The above architecture would integrate into this architecture in such that individual RAN slices would make use of the flexible RAN. I nthe following, we elaborate on some flexible RAN technologies which integrate well into the above RAN slicing concept.

Multi-connectivity (MC): The term RAN MC refers to the versatile scenario where a UE connects to the network via multiple cells. For the sake of the current explanation, it suffices to consider that a multi-connectivity approach takes place whenever the connection of the UE to the RAN involves multiple PHY interfaces. Those multiple PHY interfaces are leveraged to deliver enhanced performance capabilities, which are translated into aggregated throughput or increased reliability. A major challenge is to enforce different QoS requirements, differentiation, and prioritisation within a RAN exploiting MC and Multi-RAT through a single scheduler.

Next, we consider two MC options, namely the *common PDCP* and *common MAC approach* [R+16]. The *common PDCP approach* dictates that the PDCP layer of the protocol stack is shared between the individual connections of the RAN multi-connectivity (henceforth called "radio leg"), and all layers below PDCP are separate logical entities. The main advantage of the common PDCP approach is the flexibility it offers in terms of the physical location of the protocol stack layers. In particular, since the interface between PDCP and RLC is not a time-critical interface, the common PDCP layer is not necessarily co-located with RLC, hence mobility-related signaling can be hidden from the CN. In the *common MAC approach*, the multi-connectivity anchor point is the MAC layer. Owing to the time-critical interface between MAC and PHY, the common MAC approach requires that either the multi-connectivity legs originate from the same site, or they are interconnected via a high-capacity transport link. Nevertheless, the common MAC approach offers the advantage of fast information exchange between the different multi-connectivity legs. This facilitates coordinated scheduling, interference mitigation, and other schemes related to MAC scheduling such that the overall performance of the radio network is improved.

Figure 3-6 illustrates the integration of RAN slicing Option 1 and both multi-connectivity options explained above. In both cases, the implementation of multi-connectivity would be customized for each slice such that each tenant may use its own preferred multi-connectivity option. On lower PHY layer (PHY-TP), the SDM-X would then coordinate the access to the individual RATs or RAPs (again Network Control Layer is not shown here; c.f. Figure 3-2). Furthermore, both multi-connectivity options are symmetric with respect to the common network functions such that the actual multi-connectivity choice would be transparent to the underlying common network functions.

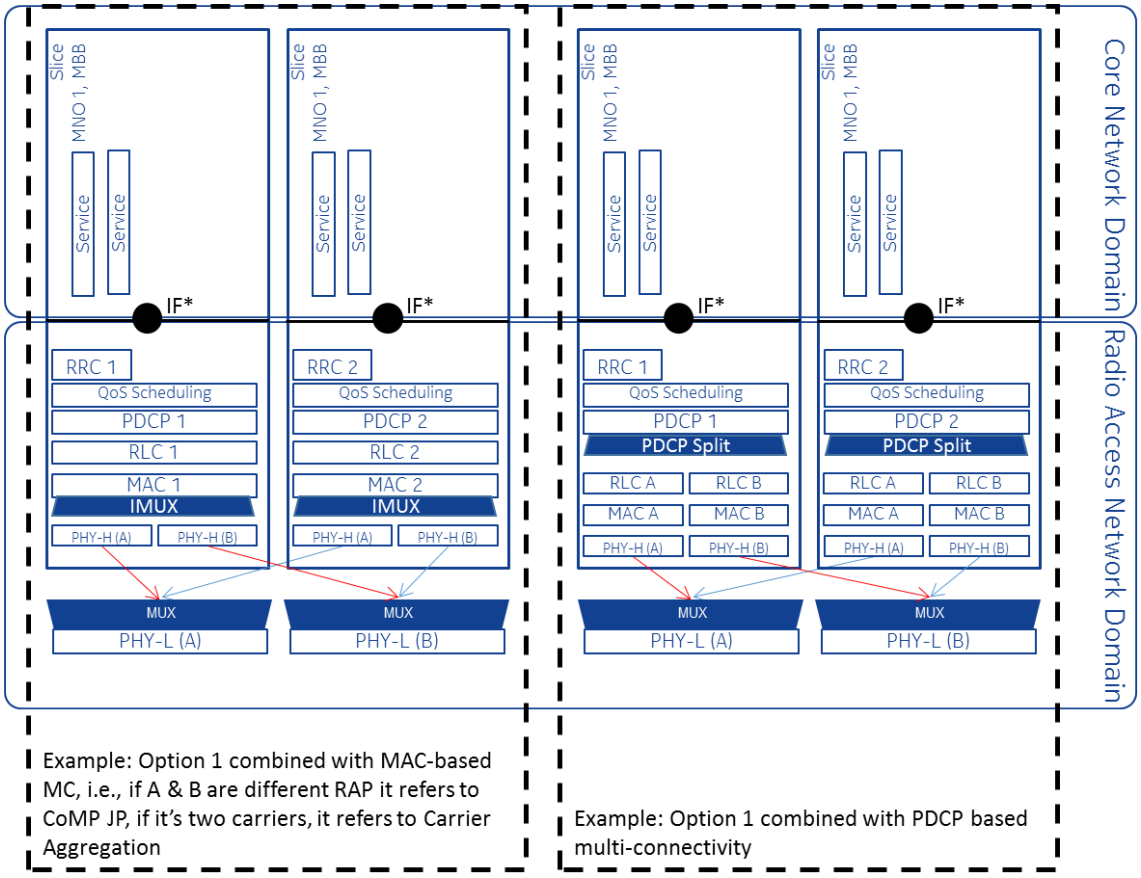


Figure 3-6: Integration of multi-connectivity with RAN slicing Option 1

By contrast to the previous example, Figure 3-7 illustrates the integration of RAN Slicing Option 2 with both multi-connectivity options. In the case of MAC-layer multi-connectivity, the operation of multi-connectivity would be transparent to the individual slices because multi-connectivity would be implemented using the common network functions. The SDM-X would then control the access of the individual slices towards the common multi-connectivity setup. In the case of PDCP-layer multi-connectivity, each slice already needs to split its service flows into individual sub-flows which are then delivered through the individual RATs or RAPs. The advantage is that the tenant of each slice may control the distribution of traffic towards the individual multi-connectivity RATs/RAPs. On the other hand, each slice must be aware of the available multi-connectivity options, i.e., which is controlled by the SDM-X.

Multi-RAT and millimeter wave (mmW) technology: It is envisioned that mmW technology will play a key role in the fulfillment of 5G network requirements, in particular for mobile broadband services. MC will be essentially required to support mmW deployments. The architecture for these deployments will depend mainly upon backhaul capabilities, cloud implementations (edge or core cloud), and the availability of mmW Remote Radio Head (mmRRH). Consequently, a flexible architecture incorporating mmW support is required to meet different slice requirements. A MC architecture based on stand-alone mmW deployment will be sufficient for a slice hosting broadband services. By contrast, a slice hosting critical services should provision the MC architecture including sub-6 GHz nodes for the assistance of mmW access.

Mobile edge computing and edge cloud processing: Advanced 5G services are envisioned to be offered at the network edge so as to reside much closer to the user in order to enhance delay

and perceived performance, e.g., adopting the ETSI MEC paradigm[2]. Therefore, a flexible service chaining should also be improved to establish dynamic services considering edge network locations and might be combined with VNFs to ensure a joint optimisation of services and networking operations. Edge server locations can also be exploited for storage, computation and dynamic service creation within a given network slice by verticals and OTT providers, introducing another multi-tenancy dimension.

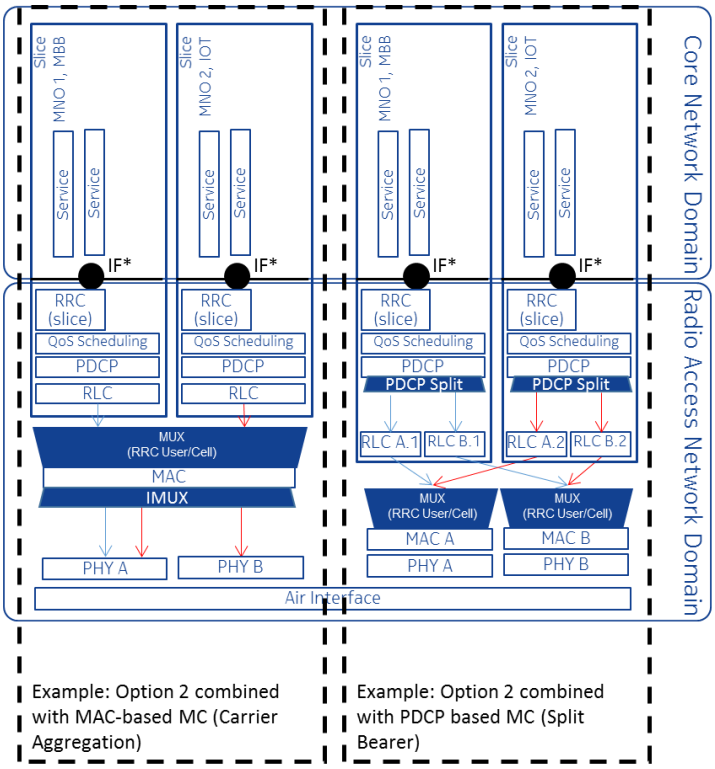


Figure 3-7: Integration of multi-connectivity with RAN slicing Option 2.

[2] http://www.etsi.org/technologies-clusters/technologies/mobile-edge-computing

4. Multi-Technology Architecture in Heterogeneous Networks

The objective of this chapter is to describe the necessary mechanisms for global and joint management of resources of multiple connectivity layers, i.e., multiple network layers such as macro and small cells, and multiple RAT layers such as below 6 GHz and mm-wave. Mechanisms include radio resource and connection management and multi-tenancy. While focusing on a heterogeneous network, one of the goals will be to design all the functions in a way that they can be flexibly placed and combined following the concept of 5G NORMA of flexible placement and allocation of functions. Finally, particular attention is payed to the orchestration and management of the functionality and the required interfaces (input/output data), which are necessary for the 5G NORMA orchestration and management framework.

4.1. Multi-Connectivity Functional Architectures

One of the key objectives for mobile networks is to provide an excellent end-user experience to satisfy the ever-growing demand on data rates which is roughly doubling every year. But the need for more capacity is just one driver for mobile networks to evolve towards 5G. In fact, 5G networks are envisioned to be unified platforms for all types of spectrum and bands, from low bands below 6 GHz to emerging higher bands such as above 30 GHz (mm-wave).

Multi-connectivity of single user terminals to multiple radio access points is a 5G key enabler in order to satisfy the demanding requirements of 5G mobile networks. Multi-connectivity supports simultaneous connectivity and aggregation across different technologies such as 5G, 4G (3GPP LTE [36.300]), and unlicensed technologies such as Wi-Fi (IEEE 802.11 [802.11ac]). In addition, it may connect to multiple network layers such as macro and small cells and multiple radio access technology (RAT) layers such as below 6 GHz and mm-wave. The latter example particularly results in improving the capacity as well as the reliability. In addition, multi-access 5G core networks will ensure mobile operators can continue to leverage today's investments. To accomplish this, 5G systems will need to support end-to-end network architectures and protocols that seamlessly combine multiple RATs and network layers together into a single virtual radio access network (RAN).

In this chapter, we present three specific architecture options, which allow for integrating multi-connectivity into 5G networks. We highlight changes applied to 3GPP LTE as well as novel functionality not yet considered by 3GPP LTE.

4.1.1. Common TCP/IP Solutions

Multi-homing, multi-path and multi-connectivity mechanisms can take advantage of a single network node with multiple network interfaces and configure the network node with multiple routing addresses (IP addresses). These mechanisms can assist the network to increase the reliability, gain, throughput, and goodput, reduce the fault tolerance, and eliminate the single point of failure. Multi-homing mechanisms are divided into two types: asymmetric multi-homing and symmetric multi-homing. Asymmetric multi-homing is the case where only one of the two end-points is able to transmit or receive from a single application address (port number) with multiple routing addresses. Symmetric multi-homing is the case where both end points do support multi-homing, and are able to transmit and receive from a single application address (port number) with multiple routing addresses at both ends. Multi-path mechanisms provide the ability to simultaneously use multiple paths between nodes, and create multiple TCP/IP sessions.

The Stream Control Transmission Protocol (SCTP) [OY02] supports multi-homing for providing network fault tolerance, network load sharing, and multiple path capabilities for transmitting user messages. Multi-Path Transmission Control Protocol (MPTCP) [FRH+13] is an extension of TCP that allows a client to establish multiple links over different network interfaces to the same network destination. It also provides multiple TCP flows across disjoint paths. A key aspect of multi-

homing, multipath and multi-connectivity is dealing with the range and variabilities of performance that can be experienced over different available link and connectivity options. Efficiency can be significantly reduced by the need to cope with the fluctuations in capability over different links, and to direct packets accordingly among the links, i.e., decide on and dynamically vary which packets should be sent on which links. One loose analogy to the issues experienced here is, for example, apparent in multi-part download managers. Download managers may waste a significant amount of time by waiting for packets on a remaining given block to arrive. If those packets were sent again on a new connection, the download would be completed almost immediately. Moreover, such an issue is very apparent if multicast and particularly broadcast solutions are employed. In this case, limited feedback on the success of packets at receivers may be available. Hence, multicast and broadcast packet retransmissions due to lost packets may not be useful for a large portion of receivers.

In such cases, a means is needed to transmit packets that are guaranteed to be always useful at receivers, no matter which packet is sent. A solution to this can be the implementation of a rateless fountain coding solution at packet-level on a downloaded file, such as RaptorQ coding [LSW+11]. Coded packets can be created almost unlimited on-the-fly and sent over the links as needed. This maximises the utilisation efficiency on each link because each packet would be useful to reconstruct the download. Furthermore, the success of decoding is very high (one chance in a million of failure) if only two more coded symbols (or packets) than the number of symbols in the download is received, i.e., very low transmission overhead. Such RaptorQ coding can be easily implemented as a sub-layer of the transport layer or application layer, residing between the conventional transport (e.g., TCP) and application (e.g., HTTP).

For a multi-connectivity case where unicast connections are being combined with a broadcast connection, Figure 4-1 presents the mapping of fountain-coded packet set for transmission over different interfaces.

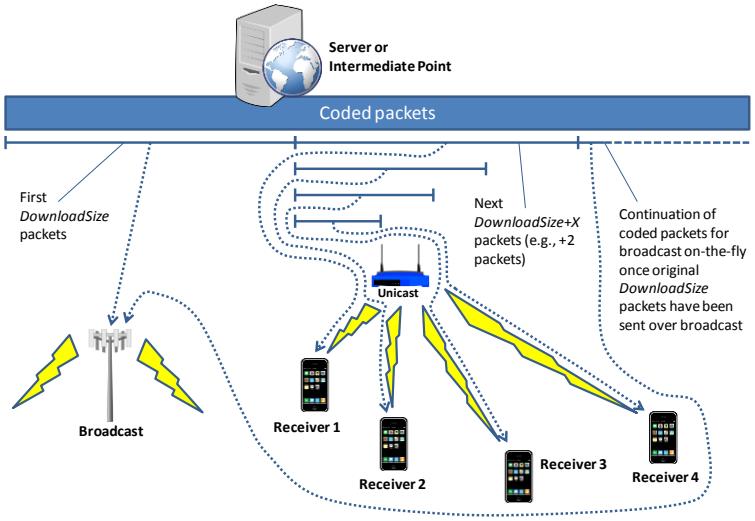


Figure 4-1: Mapping of fountain-coded packets to broadcast and unicast (non-systematic coding case) for different receivers of a software download

4.1.2. Common PDCP Solution

Multi-RAT Support (mmW)

We envision a key role of mm-wave technology in the development of 5G access networks [NGMN15]. Future deployments of mm-wave access points (mmAPs) in 5G access network will ensure the delivery of high data rates to the UEs. However, due to the special propagation characteristics in the mm-wave band, it is challenging to provide highly reliable and uninterrupted data transfer to the UEs using the mm-wave technology, especially for mobile UEs (Figure 4-2).

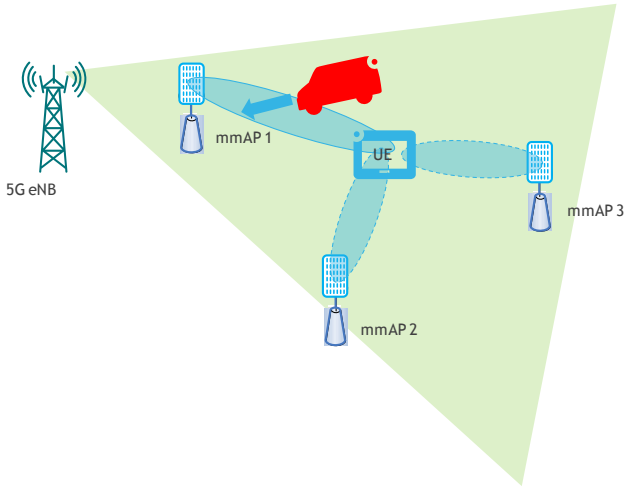


Figure 4-2: Redundant coverage for mm-wave base stations shown here as mmAP

The urban-micro mm-wave channel, as considered for our architecture, is characterised by a low number of possible paths (LOS and NLOS) between base station and UE, from which in most cases probably only one path will be used for transmission with high gain, narrow half-power beam width antenna beams. This makes transmission quality sensitive to blocking effects caused by sudden user movement or obstacles entering the transmission path, leading to poor reliability. Therefore, to minimise interruption times or ideally even to avoid interruptions and to guarantee reliability, we propose:

a) The mmAP deployments must be supported by the low-band 5G coverage layer.
b) Redundant coverage of mmAPs should be provisioned for the UEs.

For this purpose, we foresee that multi-connectivity (MC) will be an essential or rather a fundamental feature in a 5G-access network. Moreover, a UE must be able to detect and receive from multiple mmAPs to ensure the possibility of MC, link monitoring, and fast selection. To provide redundant coverage, multiple mmAPs are placed within the low-band 5G coverage area, building a "serving cluster," so that the UEs are within transmission range of each mmAP of the cluster (see Figure 4-2). It is assumed that a UE is served by at least one of the mmAPs out of the serving cluster at a time (in the example mmAP1). If the connection to the serving mmAP is blocked by an obstacle, the UE possibly will be instructed to connect to another mmAP serving the area from another direction, so that the transmission is no longer affected by the obstacle, i.e., there is no interruption in the data transfer, e.g. mmAP2 or mmAP3 can take over. It is assumed that such a cluster of mmAPs is within the coverage area of a 5G eNodeB, and the mmAPs are using the same high carrier frequency and bandwidth. However, for full flexibility, the mmAPs in a cluster may belong to different eNodeBs. This requires an efficient multi-connectivity based architecture for the 5G access network that will support intelligent radio resource management, data forwarding, and data buffering for services requiring mm-wave transmission.

In line with these requirements, we present and discuss the efficient methods for mmAP detection, cluster configuration, required functionalities, protocol mechanisms and the architecture solution to enable reliable high rate data transmission with mm-wave technology.

Detection of mmAPs and configuration of clusters*:* In LTE, neighbour cell detection by UEs is based on primary and secondary synchronisation channels. Currently a neighbour cell detection mechanism for mmWave is not specified in 3GPP. In LTE, Common Reference Signals (CRS) (pilots) are used for channel estimation and mobility handling. However, common pilots for mm-wave detection will drastically reduce the coverage of the mm-wave access point [BHR+14]. Therefore, we propose that precoded pilots should be used for mmAP detection by the UEs. In addition to that, mmAPs should be deployed in clusters because of the special propagation characteristics of mmW to enable a fast switch between mmAPs, Therefore, the UE specific cluster definition together with the coordinated pilot transmission pattern need to be coordinated and

communicated to the UEs. For this purpose, initial access schemes supported by low-band 5G nodes are required to be specified for mmAP systems supporting high gain beamforming antenna configurations.

We propose a two-step scheme where in a first step a certain degree of information about UE location within the low-band coverage area is provided to mmAPs. In addition, low-band 5G configures the UE with the pre-coded pilot structure of these mmAPs. As a second step, the mmAPs can transmit long range narrow beams using precoded pilots, which can be efficiently detected by the UE.

In case of a cluster, coordinated pilot transmission by the mmAPs in the cluster supported by the low-band 5G (or another functional unit depending on the architecture approach) will enable a UE to measure the mmAPs in the cluster. In case of mobility, new mmAPs can be configured and the cluster can be updated. For these requirements, a low band 5G node should control the UE, i.e., new RRC functionality or new RRC protocol elements:

- Definition of a mm-wave cluster (the set of possible mmAPs for each UE) by the 5G node and configuration of this cluster towards the UE.
- Definition of time, frequency and pilot sequence of precoded pilots: info to mmAPs and UEs.
- Update of mm-wave clusters in case of UE mobility.
- UE measurement configuration and UE measurement evaluation for mm-wave.
- UE-centric ID given by low-band 5G node, also valid inside the cluster.

Architecture option supporting mm-wave data: LTE dual connectivity is mainly designed for non-ideal backhaul transporting data of the X2 interface [36.842]. Therefore, using dual connectivity option 3C (split within PDCP layer), all data is processed and stored within the MeNodeB (macro cell). The 5G node supporting multiple mmAPs, as shown in Figure 4-3, would require large storage capacity and many high speed links to the mmAPs. In this case, the dual connectivity architecture, defined in LTE Release 12, will not be efficient and a deployment might even not be possible considering the high backhaul capacity demand in the mmW band.

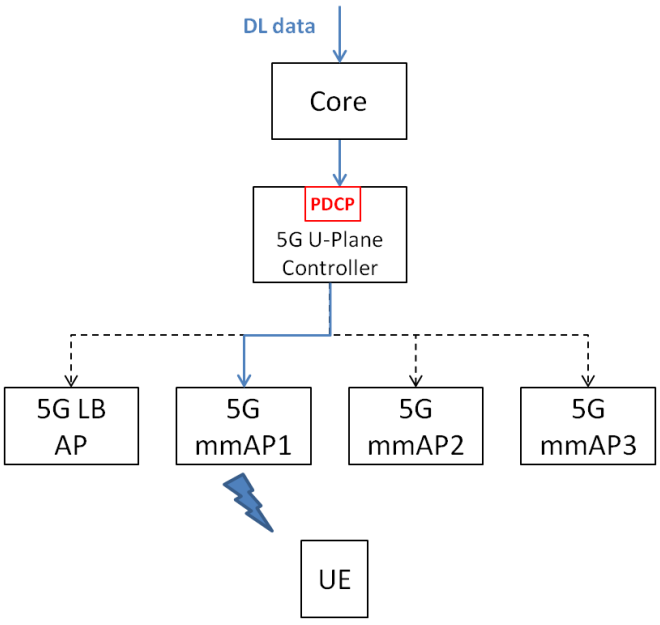


Figure 4-3: Proposed architecture solution for low band supported mm-wave 5G access network including 5G U-Plane (aka d-layer) Controller

Therefore, we propose that within the 5G access network, the data storage and forwarding functionality should be revisited. Especially, a PDCP storage and traffic steering functionality should

be defined, which forwards the data to the serving mmAP or serving cluster. However, while the configuration of clusters and mmAPs can still be controlled by the 5G control node, an additional node or access cloud functionality, which comprises the PDCP layer, should be defined, i.e. a storage and switching functionality. This will efficiently manage the data transfer during handovers between mmAPs and 5G control nodes.

In the following, we present two architecture solutions for this node or access cloud functionality, both are supported by the low-band 5G node. Figure 4-3 depicts a solution in which all UE traffic (low and high data rates) is handled by a 5G U-Plane Controller. In case a UE is moving outside the mm-wave cluster and can no longer be served by the any mmAP, low-band 5G node can still be used as fall back solution for the continuous transmission but with lower data rate.

Under the consideration of backhaul capabilities in current and future low-band radio access nodes, we propose a further split between High and Low Data Rate PDCP. For this reason, we introduce a new entity or a function in the edge cloud, which we call Radio Network Data Distributor (RNDD) as illustrated in Figure 4-4. We split the data layer in high data rate and low data rate flows in accordance with that, and we define PDCP-H and PDCP-L respectively for high and low data rates. The 5G-LB AP hosts the PDCP-L and RRC whereas the RNDD will host the PDCP-H. The 5G-LB AP, which is the RRC-Host, will manage the mm-wave cluster via RNDD. Being the RRC-host, it will also control the traffic steering in RNDD, which is hosting PDCP-H. This will require new and standardised interfaces between 5G-LB node and RNDD and it enables an optimised split between the high data rate and low data rate applications, i.e., it relaxes the backhauling requirements with respect to 5G low band access points.

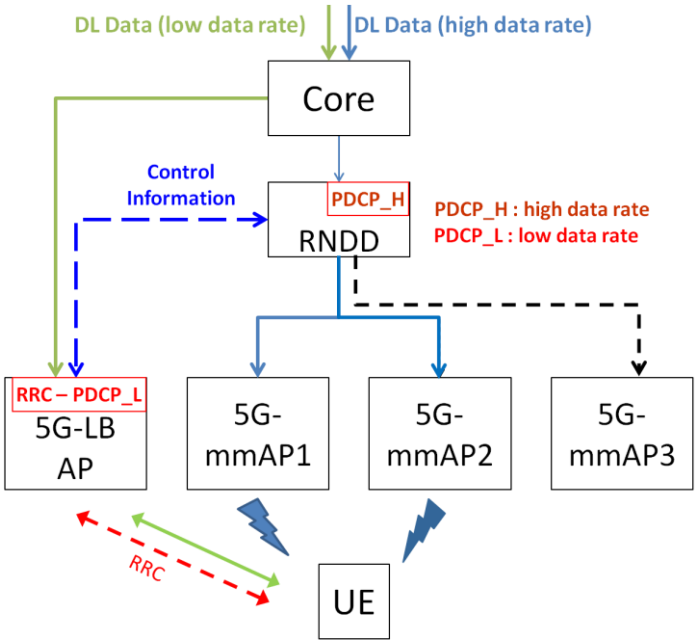


Figure 4-4: Proposed architecture solution for a split between high and low data rate PDCP

Data and Control Layer Evolution

Despite the advantages of dual connectivity in terms of throughput increase, the underlying LTE architecture is not suitable for supporting multi-connectivity as a means to address the requirements set for 5G. The main shortcomings are two-fold: i) An increased signalling overhead is associated with frequent mobility events within HetNet deployments; ii) Ultra-reliable applications cannot be supported. Next, we elaborate on the above shortcomings, and provide a solution aimed to address the above points.

Signalling overhead due to mobility. 5G network topologies are anticipated to deploy several clusters of 5G small cells with overlapping coverage area with that of a (either 5G or legacy LTE) macro cell [NOK-WPb]. Although not clearly defined yet, the number of 5G small cells within one cluster is expected to be large, i.e., some tens of small cells per cluster, owing to their limited coverage area. The limited coverage area of small cells is associated with an increased occurrence of mobility events such as handovers and cell measurements, particularly for fast moving UEs. The frequent occurrence of mobility events entails a huge signalling overhead to the RAN, involving a set of control signals associated with handover commands, which are exchanged between eNodeBs. Additionally, the current RAN architecture imply that frequent mobility events affect the core network as well because each time a handover is triggered by the RAN the core network has to switch the transmission path accordingly.

Support of Ultra-high reliability. Dual connectivity in LTE focuses on increasing the throughput by establishing dual bearer connection to the UE. In some cases, bearer split is also supported, in the sense that the UE is able to split its bearer connection to two eNodeBs, aggregating thus its throughput. Nonetheless, in LTE, ultra-reliability scenarios were not addressed, i.e., scenarios where high reliability is more critical than high throughput. Ultra-reliable applications involve the duplication of one or more bearers across multiple eNodeBs, exploiting thus the concept of diversity. On the basis of the LTE RAN architecture, a bearer duplication would involve new features which would also increase the complexity of the corresponding deployment.

The proposed architecture involves the use of an edge cloud, where the RRC (control) and the PDCP layer will be located. The remaining protocol stacks will remain on the eNodeB site, as shown in Figure 4-5. With respect to the multi-connectivity related shortcomings of the LTE architecture, the proposed architecture offers the following anticipated advantages:

a) The frequent mobility between small cells is hidden to the core network. This is because from the core network's perspective no path switch occurs each time a handover between two small cells takes place. In addition, the RRC entity in the RAN that anchors the UE mobility remains the same. This results in a considerably lower signalling overhead.
b) Data duplication across cells is facilitated: The PDCP layer in the edge cloud would be responsible for duplicating the data across multiple cells. Such feature can be more easily supported with the introduction of the edge cloud, resulting in much lower burden compared to duplication from the core network.

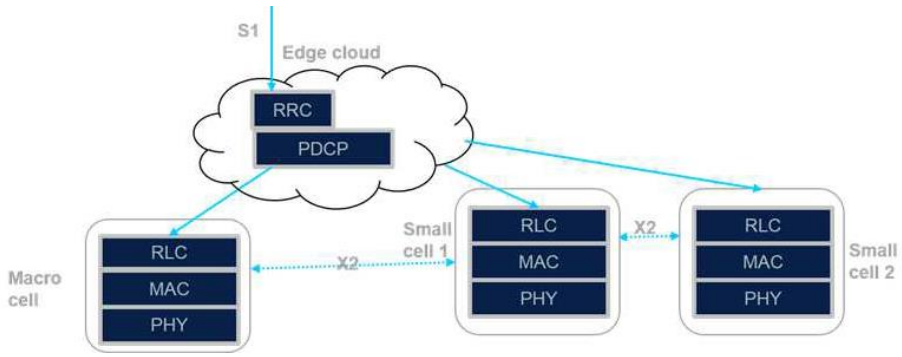


Figure 4-5: Moving RRC (c-layer) and PDCP (d-layer) to the cloud

Inter-RAT Connectivity

Inter-RAT multi-connectivity is a feature that enables the UE to simultaneously connect to more than one RAT. A multi-connectivity approach is proposed in LTE Release 12, with the launch of dual connectivity. Dual connectivity allows the UE to connect to two base stations that operate on different frequencies. The base stations are connected via the X2 interface, hence enabling direct flow of packets through split bearer. The dual connectivity approach enhances reliability of the data flow. However, it does not address the scenario of two base stations belonging to

different RATs. As LTE is a widely accepted and heavily deployed technology, the transition from LTE to 5G is critical, and will take some time. Therefore, it is of high importance to consider backward compatibility of 5G with previous standards such as LTE.

As shown in Figure 4-6, the UE is connected to multiple RATs. The RAN control functions for all the RATs are implemented in the edge cloud. These functions along with the interworking function provide integration of multiple RATs. Since the interworking of LTE with previous standards is not tightly coupled, a significant delay is introduced [ALU09]. Therefore, similar mechanisms cannot be adopted for the interworking of 5G with previous standards, due to ultra-low latency and high reliability requirements of 5G services.

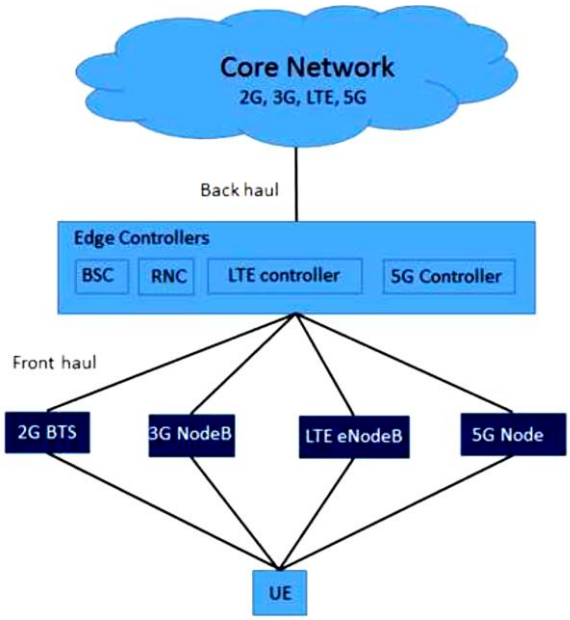


Figure 4-6: RAT multi-connectivity to UE

To provide a tight integration between multiple RATs, we propose interface Xn between 5G base stations and LTE eNodeB as shown in Figure 4-7. The introduction of this new interface will enable direct communication between LTE and 5G base stations, reducing signalling overhead by using a common control layer for both RATs, while simultaneously exploiting control layer diversity.

According to recent research [SMR+15], tight integration between LTE and 5G can be provided by using common protocol layers across RATs. Also, it is important to consider previous standards (2G and 3G) along with the tight integration of evolved LTE and 5G RATs. As shown in Figure 4-7, the higher protocol layers are common across LTE and 5G, but not with 2G/3G RAT, as the use of the same common protocol stack between 2G/3G and 5G will lead to high costs in comparison to achievable gains. The integration of 2G/3G and LTE is already state of the art and is realised via an interworking function [29.305]. We propose moving the LTE-(2G/3G) interworking function into the edge cloud and enhancing its functionalities to incorporate the interworking with tightly integrated LTE and 5G. The interworking function will run in parallel with inter-RAT mobility anchor functions located in the edge cloud. Moving the anchor point close to the edge will provide low latency handovers between multiple RATs.

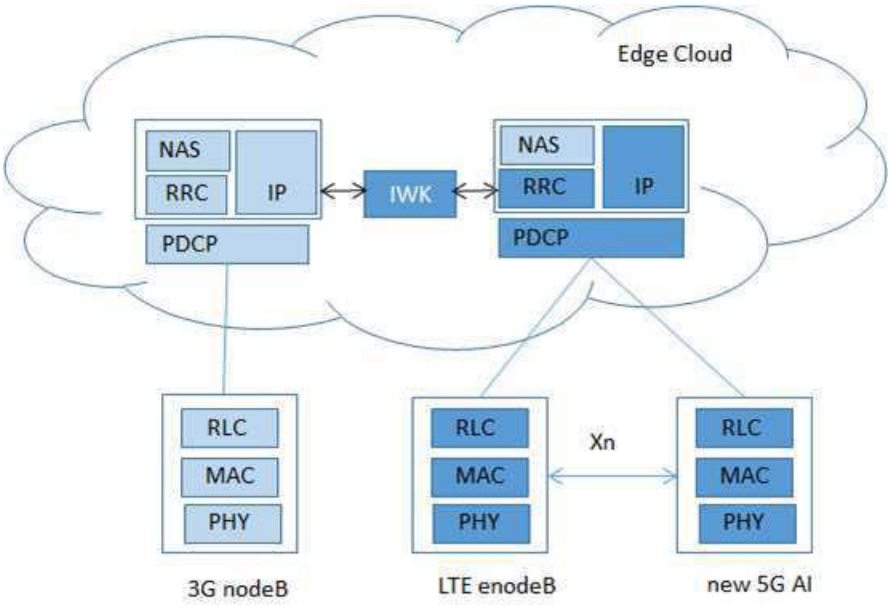


Figure 4-7: Integration of multiple RAT in edge cloud

In this section, we identify the required functionality for integration of LTE and 5G at RRC and PDCP layer. The integration of RRC and PDCP is much more feasible as the functions are asynchronous with respect to the transmission time interval (TTI) [SMR+15]. LTE and 5G are assumed to have common control and data layer. As proposed in [SMR+15], the integration can be carried out in two operating modes: diversity mode and reliability mode. The signalling and data flow in reliability mode is carried out by duplicating it over multiple RATs, increasing the reliability, and hence, no inter-RAT handover is required. However, in diversity mode, signalling is carried by either one of the RATs. Therefore, handover procedures are required and initiated if the user moves into a cell that belongs to a different RAT. We also propose dynamic selection of operating modes by RRC, depending on the QoS requirements from the UE. Reliability mode is selected if the UE requests very high data rate services such as online gaming and video streaming. On the other hand, the diversity mode can be selected in the case of high latency requirements or increase in number of users in the given area. The data is then transmitted simultaneously through multiple RATs without duplicating it to serve the hard latency constraints.

To provide close integration of RATs, we follow the architecture shown in Figure 4-7. We identified new functionalities that are required to provide inter-RAT multi-connectivity:

- *RAT selection:* This function enables the selection of a RAT depending on the measurement reports of the neighbouring cell and RAT from the UE, and its QoS requirements. RAT selection function selects either one or multiple RATs to provide data flow and signalling to a single UE. The RAT selection function operates closely with the QoS and the inter-RAT traffic management functions.
- *Operating mode selection:* This function selects the operating mode, either reliability mode or diversity mode. Different modes can be selected for control layer and data layer. For instance, control layer operates on diversity mode, allowing signalling messages to be transmitted over multiple RATs without duplicating, while data layer can operate in reliability mode, allowing duplication of data flow via multiple RATs, and vice versa.
- *Inter-RAT traffic management:* The traffic management function operates on the network layer. It manages the network load across different RATs and provides inputs to the RAT selection function.
- *Control and data flow routing:* The function is responsible for the routing of data packets and control packets if reliability mode is selected. Duplicated packets are routed through multiple RATs, increasing the throughput, or different data packets are routed through different RATs to satisfy low delay requirements.

- *PDCP sequence number synchronisation:* The common PDCP layer across multiple RATs requires additional functionality to synchronise sequence numbers of PDUs across multiple RATs. The sequence numbers of the PDUs need to be adapted, with the addition/release of RATs connected to the UE.
- *Multi-RAT support for RRC and NAS:* Additional functions like connection establishment, modification, and release for multiple RATs, or service flow mapping to RBs for multiple RATs are required.

Additional functionalities such as inter-RAT mobility management, resource scheduling, and interference cancellation are also necessary to achieve inter-RAT coordination gains.

4.1.3. Common MAC Solution

For scenarios where multiple legs of a multi-connectivity connection originate from the same physical site, the common Medium Access Control (MAC) case is envisioned. Common MAC refers to the case where the multi-connectivity legs share the PDCP, RLC, and MAC layers of the protocol stack, in a way similar to carrier aggregation [36.808]. An illustration of this idea is provided in Figure 4-8.

The main benefit of using the common MAC approach instead of the common PDCP approach is the faster switching between the legs. Particularly for mm-wave frequencies, where abrupt channel variations are anticipated, the common MAC approach is more robust than the common PDCP approach because the switching occurs at a lower layer in the protocol stack. On the other hand, however, the common MAC approach is limited to the collocated scenarios. Hence, the common layers of the protocol stack must be located at the same physical location, while only the physical layer is separated in different remote radio heads. In fact, it is the delay caused by the backhaul connection between different sites, which renders the separation of the MAC layers impractical for multi-connectivity implementations. The reason is that the packet segmentation carried out in the RLC layer must be able to follow the link adaptation messages coming from the MAC layer, and this is achieved only via a low latency connection.

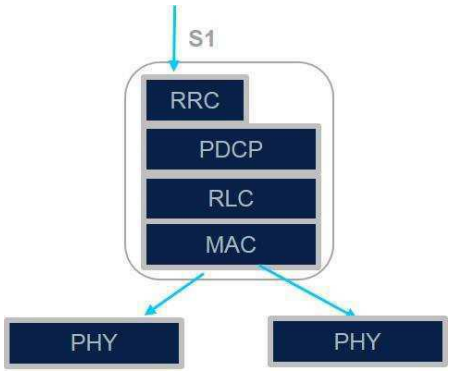


Figure 4-8: The common MAC approach

Nonetheless, it is worth pointing out that the common MAC approach does not necessarily contradict the common PDCP approach, but can rather be used on top of it. In such a case, the common PDCP layer would be located in the cloud while RLC and MAC layers are located near the antenna site. To put it in another way, the network perception of the cloud-based PDCP layer is independent of whether the common MAC approach is employed.

Table 4-1 lists the functionalities per protocol stack layer that need to be modified or fundamentally built in order to enable us to realise the novel 5G NORMA multi-connectivity architecture.

Table 4-1: List of functionalities impacted or to be built by multi-connectivity architecture

Classification	Function block (protocol layer)	LTE functions modified	New functions introduced with 5G	Contributor
Core	RAT Selection	new with 5G		Nokia, NEC
	Network Management		• Inter-RAT traffic management • Inter-RAT frequency management	TUKL
	Mobility Connection Management		• Inter-RAT mobility management	Nokia, TUKL
RAN	RRC		• Control flow routing • Mm-wave functionality (ALU)	TUKL, Nokia, ALU
	PDCP	• PDCP split bearer • Data transfer • Routing • Timing • Reordering	• Mapping between service flow and radio-bearer service for enhanced QoS support and in-service-flow differentiation • Routing and flow control for enhanced RAN level multi-connectivity with possible PDCP level radio bearer split and cloud-RAN support • Multiple flow across multiple/single RAT(s) • Single flow across multiple RATs • Further anchoring functions for flexible, on-demand d-layer enhancements, including security and mobility on demand, c/d-layer separation and cloud-RAN support • PDCP Storage Functionality (ALU)	TUKL, ALU, Nokia
	RLC	• Buffering and transferring of PDCP PDUs • Reordering and duplicate detection of RLC PDUs • Reassembly of RLC SDUs • Re-segmentation of RLC data PDUs		ALU, Nokia, TUKL
	MAC Scheduling		• Scheduling info exchange • Common priority handling • 1x uplink coordination, • UE radio network identities • Inter-RAT resource scheduling • Inter-RAT interference cancellation (IRI)	Nokia, Nomor
	MAC UE	• Link adaptation • HARQ	• Radio interface	TUKL, NEC
	Transport		• Higher layer coding: cross layer aspects	KCL

4.1.4. Applicability

Advantages of common PDCP

Flexibility: The main advantage of the common PDCP approach is its flexibility in terms of the physical location of the protocol stack layers. In particular, since the interface between PDCP and RLC is not time-critical, the common PDCP layer is not necessarily co-located with RLC. This enables different deployment structures such as a dedicated master node hosting the PDCP layer, a Data Distributor functionality within the access cloud, which distributes the PDCP data to individual access points, or a data distributor with one centralized scheduler, which transmits MAC packets to a selected remote radio head. Considering backhaul capabilities, different PDCP entities (PDCP-L, PDCP-H) can serve one UE with e.g. voice over IP and a parallel broadband video download.

Less Complexity: The basic concept of a common PDCP layer is characterized by asynchronous user and control layer, relaxed synchronization requirement between the access points, which are involved in the multi-connectivity setup, distributed MAC layer scheduling, and relaxed front haul. Based on the individual scheduling of the PDCP data in each access point, the implementation of the scheduler will be less complex compared to the common MAC solution. The scheduler has to process MAC packets only for one radio cell, i.e. a distributed processing load with respect to resource allocation of the air interface can be achieved.

Multi-RAT Support: Different RATs, e.g., a mmWave access points and a 5G low band access node, can be combined for a PDCP based multi-connectivity even with different TTIs within the different RATs. Additionally, the 5G-low band access node is also available as a fall-back solution for data transfer to the UE if a connection to any 5G-mmWave access points is not possible.

Hiding of UE mobility: With the definition of an anchor node or an anchor functionality inside the access cloud, which distributes the PDCP packets to a cluster of access points, the mobility of a UE inside the cluster can be hidden towards the core network. This results in a reduction of mobility based control messages both on the air interface and towards the mobility control entity inside the core network.

Adaptation to different 5G use cases: 5G NORMA WP2 has defined several use cases with different requirements for QoS. By configuration of different PDCP multi-connectivity implementations, many of the diverse QoS requirements can be fulfilled. The basic concept of a common PDCP enables a capacity increase of a mmWave high speed transmission. The concept based on synchronisation and an advanced buffer management of the PDCP layer enables parallel connections (diversity) from more than one access point. This will result in a high reliability or aggregated throughput, which are required for, e.g., IoT and V2X applications.

Disadvantages of common PDCP

The basic concept of a common PDCP approach has generally lower requirements with respect to backhauling. For specific implementations, e.g., a mmWave cluster to support a seamless high speed transmission to the UEs, all access points have to be provided with high speed links.

The implementation of a common PDCP, e.g., the parallel connections from more than one access point, requires a backhaul with low delay to enable the synchronisation of different PDCP stacks inside the access points, which are involved in the diversity transmission. This fast synchronisation comprises an information exchange to avoid duplicated transmission of PDCP packets in case a negative HARQ feedback was received by one access point while the transmission from a second access point was successful. Such a synchronisation protocol will also have impact on the buffer management inside the access points and the UE. Probably, the specification of such a synchronisation protocol has to take into account different performance classes with respect to backhauling, leading to different protocol options which have to be taken into account during the deploying of a common PDCP.

Advantages of Common MAC

Multiplexing Gains and Centralised Resource Management: The integration of different RAT at MAC layer can lead to large coordination gains. A large number of UEs can be served simultaneously if the resources across RATs are shared. The common MAC architecture will enable efficient resource management if the global view of the network resources will be available. The common MAC layer will allow dynamic multiplexing of uplink and downlink user data across multiple carriers belonging to different RATs. If the resources on a given RAT are not utilised due to low user demands, they can be shared with other RAT. Also, the underutilised AP can be switched off if the UEs from that cell can be served by other RAT's AP.

Inter-cell Interference Coordination: Common MAC will enable cross carrier aggregation techniques, i.e., data for a single UE can be sent over different carriers in order to avoid interference in case of multi-connectivity supported UEs. Also, with strategically managing the power distribution to the RATs, the frequency reuse distance can be minimised without increasing the inter cell interference.

Inter RAT flexibility: In case of the common MAC approach, the Inter-RAT handover process can be significantly faster as the decision can be made at MAC layer. There will be no need of a separate connection establishment process if all the above protocol layers between different RATs are common. The UE can switch flexibly between RATs by resource assignment of different RAT.

Disadvantages of Common MAC

Synchronisation: Synchronisation is the main challenge to have an integrated MAC layer across different RATs. Every RAT will have a different resource structure, TTI, and HARQ/ARQ feedback mechanisms. For example, the frequency spacing between LTE subcarriers is 15 kHz, which will be significantly different for 5G mmWave. In the case of multi-connectivity to a UE, the inter RAT resource scheduling and decoding at UE will be complex.

Co-located RATs: The main functions of MAC layer are user data scheduling, resource allocation, power allocation, and link adaptation. These functions are required to be executed frequently due to channel variations and impose stringent latency constraints. To satisfy these latency requirements, it is necessary for common MAC to be deployed close to the edge. Therefore, the common MAC approach might lead to the necessity that both RATs are co-located causing no gains in terms of network area coverage.

Increased Complexity: In order to achieve high synchronicity between RATs, complex scheduling algorithms need to be designed. Also, the complexity of existing scheduling algorithm will increase to multiplex the resources from multiple RATs.

4.2. Radio Resource and Connection Management in HetNet

Deployment of heterogeneous wireless access networks is inevitable due to increase in demand of multimedia applications such as video streaming and voice over IP (VoIP), each imposing strict Quality of service (QoS) requirements. The use of small cells such as relay nodes, femto, micro, and pico cells are among the most promising technologies to meet the increasing need for higher data rates. Also, considering the availability of various access technologies (WiFi, WiMAX, LTE), it is difficult for a network operator to find a reliable resource and connection management approach to select the best network, which ensures user satisfaction while simultaneously increasing the network efficiency. It is a crucial task to manage the radio resources in the presence of heterogeneous networks. An RRM framework for HetNets can be operated in two stages, i.e., information gathering and decision making [PKB+11]. In the following section, these mechanisms are described in detail.

4.2.1. Information Gathering

The information is collected either from the user, the network, or from both. The gathered information can also be classified into two types, i.e., predetermined and time varying [PKB+11]. The information that is required to establish connections is called predetermined information. It can be the user specified constraints, e.g., preference is given to WiFi over 3G mobile data. In case of heterogeneous access networks, further information such as AP load, maximum achievable data rate, the latency over that AP, or the type of traffic that can be supported is essential. It is also very important to know the UE capabilities supporting heterogeneous networks. If we consider the HetNet scenario, it will consists of several femtocells, mmWave cells, enodeBs and 5G macro base-stations together. The radio resource and connection management will hence be a complex task, as we need to gather the information not only for UEs but also for all small and macro-cells. Every network will have its own limitation in terms of bandwidth, transmit power, MCS supported and network entities involved.

Time varying information is also gathered after the connection has been established. This is the instantaneous information obtained to optimize the resource allocation of the heterogeneous access network and to maintain the established connection. The information on the current load on the AP, available bandwidth, the traffic intensity of the AP, or cell coverage are key parameters to determine the resource allocation. The radio parameters also play a critical role in the monitoring and allocation of resources. Parameters such as Signal to Interference plus Noise Ratio (SINR), received signal strength, symbol error rate, peak signal to noise ratio, and link condition are also considered as time varying parameters [PKB+11].

4.2.2. Resource Allocation and Connection Decision

The resource allocation and connection decision stage executes the decision making and the decision enforcement. In the decision making, the network selection and bandwidth allocation for the UEs are addressed. The decision enforcement makes sure that all decisions are executed. Once the decision is made, a request for corresponding connection establishment is sent to the network in the decision enforcement phase. The request can be accepted or rejected depending on the network load. The decision enforcement phase monitors until a successful connection is established and the resource and bandwidth allocation is executed accordingly.

As the decision mechanism in radio resource management plays a very important role in establishing connections and allocating resources, they are classified based on their applied approach into network centric, user centric, and collaborative approaches if the decision is made by the network, user, or both, respectively. The mechanisms can also be classified if the decisions are centralised or distributed in addition to the hybrid way. An elaborative survey on the techniques that are used for these approaches is presented in [PKB+11].

Network-centric Approach

The main focus of this approach is to maximise the utilisation of the network rather than on the user satisfaction. Several approaches have been proposed for the network centric RRM. In the following, some of these approaches are briefly described [PKB+11].

Statistical resource allocation schemes are designed to optimise the resource allocation by considering the probability that the service is rejected as well as the network utilization parameter. According to the statistical properties of the service rate, the resources are assigned to increase the network utilization efficiency. This scheme maps the RRM problem to an optimization problem by considering the service demand rate on the network and defining constraints over the blocking error probability to increase the network utilization factor in total. The optimization function then minimizes the cost occurring due to over allocation of the resources to the UE. The cost function for underutilisation of resources for every network and the gain per allocation of user to all available networks are included in the optimisation function. Imposing all this constraints, the objective function maximizes the profitability and reduces the cost.

Another widely used network centric approach is based on cooperative game theory. A game theory approach takes into account the allocated bandwidth from different networks to the user. Many game theory algorithms are used for allocating bandwidth in presence of heterogeneous networks. The most common approach is the N-person cooperative game or bankruptcy game theory. In this approach, each access network allocates a certain amount of bandwidth to the new connection request depending on the characteristic function and Shapley's constant [PKB+11]. The framework has been developed in [DH+06], which addresses the challenge of resource allocation and connection management by a bandwidth allocation algorithm and admission control algorithm. The approach is used when the total requirement for the bandwidth is more than the total available bandwidth of the heterogeneous network. Every network contributes by allocating the free available bandwidth on the network in order to increase network utilization efficiency. The connection management is based on the admission control algorithm.

Also, priority based resource allocation techniques play a significant role in heterogeneous scenarios. Many high priority services such as mission critical and public safety services will be addressed by 5G. It is important to design special resource allocation techniques for such services. In this resource allocation schemes, the resources allocated for low priority services can be utilized by high priority services. These allocations are done instantly by establishing a connection to the APs that have less low priority services but good signal strength. The authors in [XBC+05] developed a similar concept of degradation utility for the assignment of resources according to the user priority.

User-centric Approach

The user-centric approach focuses on UE gains rather than network resource optimisation. The techniques of load balancing and traffic distribution are not considered, which may lead to network congestion. The other drawback of this approach is in the case that the user request cannot be met by the network, a UE may waste energy for unsuccessfully trying to establish a connection to the network repeatedly [PKB+11].

Collaborative Approach

In the collaborative resource allocation scheme, both users as well as network operator participate in decision making. The mostly used scheme is a fuzzy logic controller. The first stage is a preselection stage that collects the requirements from the network, user, and application, and checks if the network satisfies the requirements [PKB+11]. The decision rules are defined based on variables such as network data rate and required application data rate. The complexity increases as the number of metrics increases [WLM+05]. The best illustration of this approach is dynamic QoS based resource allocation. In order to satisfy the customer, several QoS based RRM schemes are proposed. Dedicated resources are allocated in some of the proposed schemes, however, this mechanism leads to bandwidth usage inefficiency. Dynamic QoS based resource allocation is therefore proposed, which upgrades and degrades based on the availability of resources. These techniques satisfy the QoS requirements of a user while simultaneously increasing the network utilization efficiency. Another example is 'In Service Flow Differentiation' that allocates the radio resources based on the application requirements and is a currently widely accepted technique.

4.2.3. Centralised, Distributed and Hybrid RRM

Based on the location where the resource allocation is carried out, the RRM is classified into centralised, distributed, and hybrid RRM approaches. In centralised RRM, the controller is placed in the core. The controller is aware of the complete network and hence can achieve coordination gains. However, the overhead is significantly increased due to frequent transfer of information to the core. The other approach is distributed RRM in which the resource allocation is carried out locally at each access node. Distributed RRM does not add signalling overhead but also has no coordination gain because load balancing cannot be realized. Therefore, a combined approach is proposed, which consists of a centralised controller along with distributed assisting nodes. The

distributed nodes collect the information about the best networks available based on the radio conditions. The global optimisation is then done by the centrally located controller to achieve load balancing [KBT+10], [LCL+14].

4.3. HetNet Multi-Tenant Concepts

On-demand network sharing provides a new degree of flexibility for multi-tenancy systems compared to the first generation of network sharing concepts, which were based on long-term contractual agreements.

Resources are acquired on a short-term scale (minutes) leaving the actual allocations to signalling feedbacks. The synchronisation in resource sharing is guaranteed by a central resource management entity, which is represented by the capacity broker, within the MNO infrastructure. A tenant request reaches the capacity broker, which has a global view of the network resource utilisation. Based on such information, the capacity broker decides whether to accept or reject the tenant request aiming at optimizing the resource utilisation while maximizing the overall profits.

4.3.1. Resource Provisioning: Overview

We want to design a new algorithm that allows allocating the resources among tenants in a flexible way. Our algorithm takes as input:

1. a set of resources required from the different tenants,
2. the corresponding period of time,
3. their location,
4. the QoS constraints,
5. the probability of rejecting a call.

Based on this input, the algorithm allocates the available resources to optimize and maximize the resources' utilisation. With the proposed criterion, resources are only reserved for tenants that need them, and they are not reserved when they are not used. As a consequence, free (unused) resources are available for other tenants. Furthermore, the tenants do not need to request (as in 3GPP) more resources than needed, because the proposed criterion provides each tenant with the needed resources. Obviously, if the requested resources are not available, an outage will probably occur.

Additionally, the algorithm allows for different pricing (or sharing) levels according to the tenants' needs. We also take into account the number of users of each tenant and their location (where demand is higher and consequently the resources are more valuable).

4.3.2. Admission Control for Tenant Requests

The admission control is in charge of monitoring the available system capacity and leasing resources to different tenants. Two different types of traffic are considered in our mechanism: (i) Guaranteed with resources locked for explicit use of a mobile virtual network operator (MVNO) and (ii) best effort (BE) where resources are pooled and shared by all participants. The admission control performs (i) analysis and forecasting of the tenant traffic and (ii) identification of the limit to slice the available resources into these two types of traffic classes, depending on the forecasting and its associated Confidence Degree (CD). We envision a more stringent CD for stronger traffic guarantees and a looser slice limit for best effort traffic types. This encourages the capacity broker to overbook available resources for such types of traffic which may experience delay and loss.

Forecasting Techniques

Different forecasting solutions have been evaluated in our solution. Specifically, due to the flexible, on-demand nature of the tenant requests, we study methods to predict capacity on a short term basis. Short-term traffic is challenging to forecast, due to stronger variations that may be observed and have a lack of periodicity. To compensate for the inability to appropriately perceive the non-uniformities of user traffic, we analyse the traffic sub-components of the forecasting.

Specifically, we extract periodical patterns using the fast Fourier transform (FFT) and analyse the predicted traffic in the time domain by transforming it into *n* distinct components into the frequency domain. Then, we use the inverse fast Fourier transform (IFFT) to transform each periodical component into the equivalent time series, which provides a more regular traffic pattern compared to the initial prediction. In this way, we manage to decompose the prior traffic into *k* periodic components that can be predicted more accurately compared to the initial non-uniform traffic. In addition, we represent the experimentally estimated capacity values in three different CDs by using the student's distribution, over the average value of the predicted capacity across 1000 simulations. We make the assumption that the sample size of the estimated capacity is large and its standard deviation is unknown.

Preliminary Results

Here, we study the performance of our solution for varying forecasting CDs in a scenario where only VoIP requests require guaranteed capacity. The scenario is studied for the time duration of the prediction (i.e., 20 minutes), while increasing the offered load. To evaluate our solution we compared it with the baseline scenario where admission for tenant requests is based only on resource availability at the arrival moment of the request. A Monte Carlo event-based simulation is carried out in MATLAB with 1000 iterations to achieve statistical validity for each forecasting step. In Figure 4-9, we show the RB utilisation (a) without SLA violation provided that the offered load consists of requests and (b) SLA violation versus the offered load associated to the tenant requests.

When our solution is applied, the utilisation without SLA violation is increased far beyond the baseline approach as we introduce more offered load. On the one hand, in Figure 4-9(a), the curves of RB utilisation stop at a certain offered load limit beyond which we accept no more requests due to SLA violation (i.e., 9600 kbps for the baseline scenario, 20800 kbps for 90 % CD, 24000 kbps for 95 % CD and 27200 kbps for 99 % CD). The vertical dashed lines point out this limit, i.e., successfully served requests' load without violating SLAs. The maximum amount of offered traffic without an SLA violation for 90 % CD results in 23 % higher utilisation compared to the baseline scenario, whereas the utilisation percentage for 95 % CD is 17.27 %. Despite the fact that our scheme uses 99 % CD, which results into 42.3 % (i.e., lower than the 44.2 % achieved by the baseline scheme), it can accommodate higher offered load associated with MVNO requests (i.e., 27200 kbps). In addition, Figure 4-9(b) shows the SLA violation for the different approaches. At the maximum offered load of requests at which our solution with 99 % CD still shows no SLA violation (i.e., 27200 kbps) the baseline scenario already results in 68.63 % violations. For the same offered traffic, our solution with 90 % CD and with 95 % results in 52.71 % and 19.3 % SLA violations, respectively.

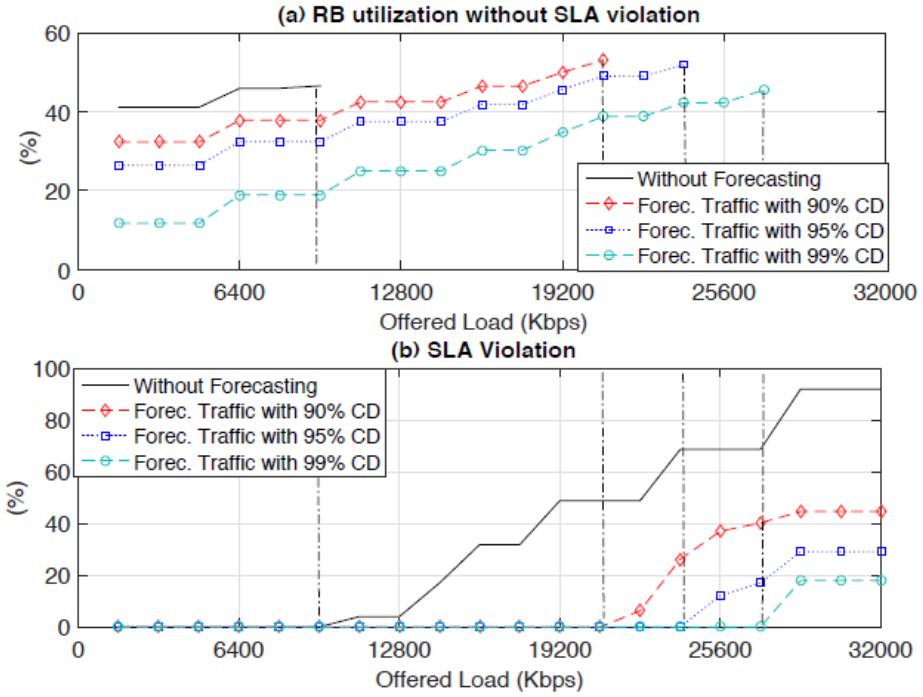


Figure 4-9: RB utilisation (a) without SLA violation and (b) SLA violation versus the offered load

4.3.3. Maximizing Tenant Revenues

The new sharing criterion provides the network provider with the possibility to maximise the resource utilisation. Furthermore, we want to design a new criterion that also allows maximizing the infrastructure provider revenue.

To achieve this goal, the problem is to find a rule that allows for deciding if it is better to accept or reject an incoming new request from a tenant in order to maximise the overall profit. The trade-off is between (i) accepting a request now, thereby obtaining immediate revenue but potentially losing the possibility to accommodate a future request with higher reward, or (ii) rejecting the request in order to keep the resources free for a possible future higher offer that the infrastructure provider may receive in the future.

Optimal Stopping Theory

In order to design an algorithm that addresses the above problem, we leverage on optimal stopping theory [CRS71]. In general, optimal stopping theory is concerned with the problem of choosing a time to take a particular action, in order to maximise an expected reward:

Given a sequence of random variables X1, X2,…, Xn (whose joint distribution is assumed to be known) with an associated reward, at each step we choose either to stop observing (that is, to accept the request) or continue (that, is, reject the request and wait for the next one). If we stop observing at step i, we will receive the corresponding reward.

Optimal stopping theory provides a stopping rule to maximise the expected reward. In particular, it gives a threshold such that we accept the first offer higher than this threshold.

In order to gain insight into the application of this theory to a multi-tenancy scenario, we have simulated a simple scenario composed of two incoming request classes (elastic and inelastic) with fixed duration and amount of capacity required, a fixed reward (different for each), and a known distribution. For this simple case, we have obtained a threshold rule using optimal stopping theory. In particular, we assume that inelastic traffic corresponds to GBR, so that for this class we

accommodate the request only if there is enough capacity available, and the elastic traffic corresponds to best effort, so that no QoS is guaranteed for this class, and their request is accommodated only if the reward is sufficiently high.

Preliminary Results

We have compared the average return per unit of time obtained with optimal stopping theory with the following two alternatives (note that inelastic traffic is always accepted in the optimal algorithm):

1. We never accept elastic traffic;
2. We always accept elastic traffic.

In Figure 4-10 below, we show the average return per unit of time obtained as a function of the request duration b for a ratio between elastic and inelastic reward equal to 0.3.

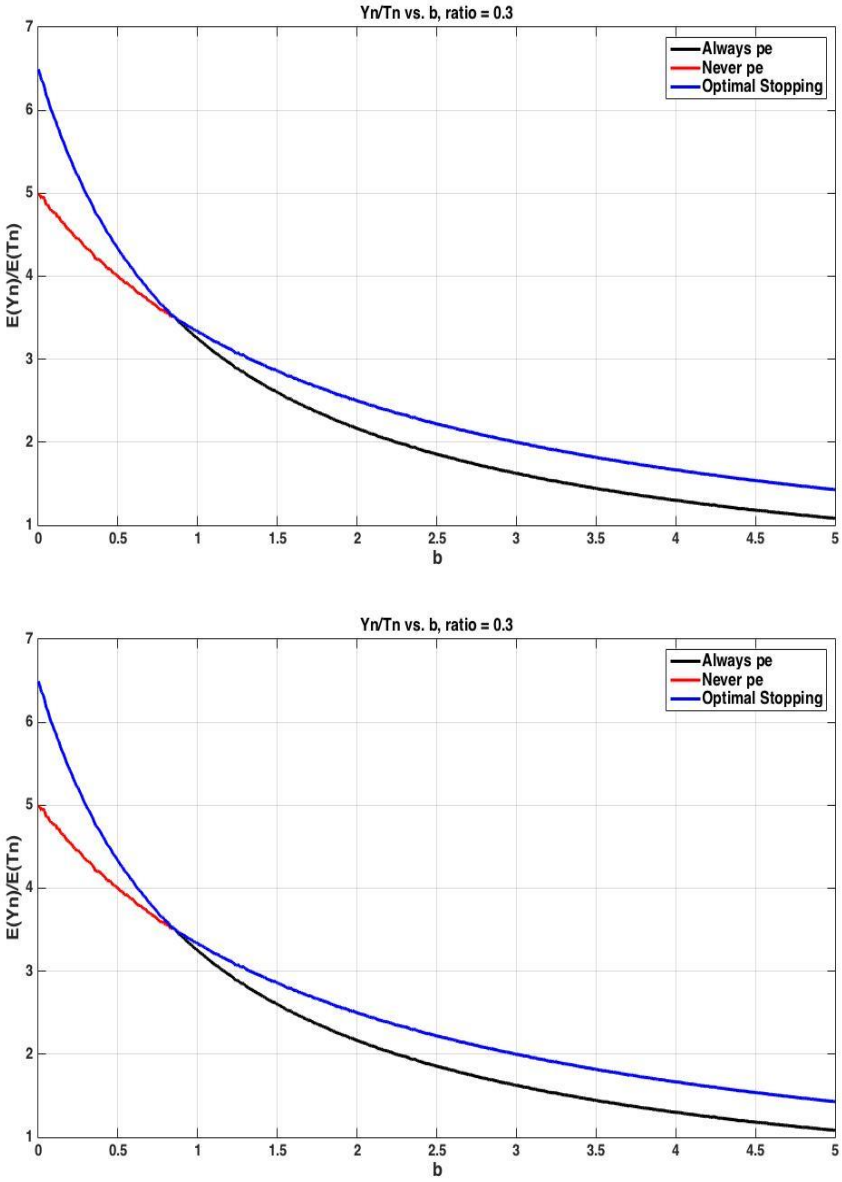


Figure 4-10: Average return (per time unit) versus request duration

As shown, optimal stopping theory (blue line) always provides the maximum revenue, which confirms the suitability of this tool for this purpose. Specifically, we can conclude that with optimal stopping theory, the revenue in a simple scenario can be maximised resulting in a threshold that allows us to decide whether to accept or reject the incoming requests.

In a more real scenario, an infrastructure provider receives multiple requests from tenants characterised by:

- A certain amount of resources to be reserved;
- The starting and end times for the reservation;
- The bid offered;
- Required quality and SLA (e.g., elastic or inelastic traffic).

Based on the above parameters and the amount of available resources, in order to maximise his revenue, the infrastructure provider decides to reject and accept the requests. Using Optimal Stopping Theory, it is challenging to model all these parameters , which motivates the following approach.

Semi-Markov Decision processes (SMDP) and Q-learning

Markov decision process (MDP) is a versatile and powerful tool for analysing probabilistic sequential decision processes when outcomes are uncertain. The MDP consists of decision epochs, states, actions, rewards, and transition probabilities. Choosing an action in a state generates a reward and determines the state at the next decision epoch through a transition probability function. Policies or strategies are prescription of which action to choose under any eventuality at every future decision epoch. The goal is to seek the policy that is optimal in some sense. MDP provides us a powerful tool, but it works only for discrete-time systems and in a real scenario, the assumption that the offers arrive in fixed epochs is too strong.

SMDP models the infrastructure sharing as a Markovian chain characterised by states, a set of actions, rewards and transition probabilities but unlike MDP, the decision epochs are random, so it allows to model continuous-time system.

Based on SMDP and applying decision theory, we have developed an algorithm that finds the decision policy that maximises InP revenue while satisfying the service guarantees required. SMDP guarantees optimal performance but it implies two limitations:

1. Very high computational cost (as the state space is large);
2. It requires the complete knowledge of request statistics (inter-arrival and departure time) and system transition probabilities.

Due to above reasons, we have developed a new adaptive algorithm, which leverages on reinforcement techniques and allows us to obtain performance close to optimal.

In particular, our algorithm leverages on Q-learning: it is built on SMDP techniques, but instead of requiring full knowledge of the system's parameters, it learns the system behaviour by taking non-optimal decisions during a so called *learning phase*. Hence, it is an online tool that after a training period is able to find the decision policy that maximises the InP revenue evaluating the best possible action in every system state.

New simulation results

In order to evaluate the adaptive algorithm performance, we consider a scenario with two classes of incoming request:

1. Inelastic traffic demanding a certain fixed throughput which needs to be always satisfied with a fixed outage probability;
2. Elastic traffic requiring an average throughput guarantee.

In Figure 4-11, we compare the return obtained with Q-learning changing the ratio among inelastic and elastic per unit of time bid with:

1. The optimal algorithm based on SMDP,

2. Always accept all the requests;
3. Always reject elastic requests.

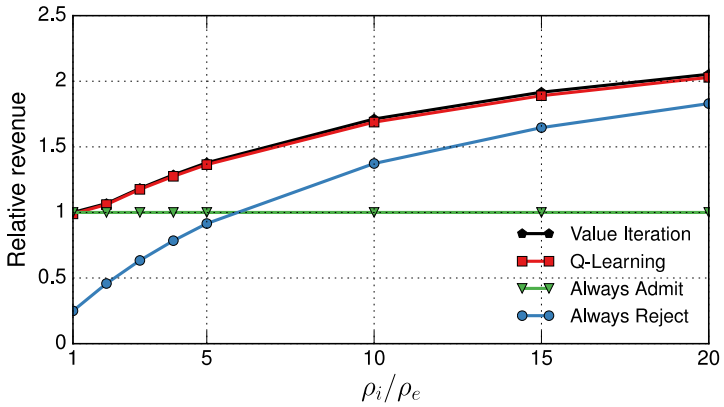


Figure 4-11: Relative return versus ρ_i/ρ_e

As shown, the optimal admission control algorithm based on the SMDP method always converges to the optimal policy, which provides the highest revenue. The adaptive algorithm designed provides close to optimal performance flexibly adapting to varying system conditions. Both algorithms provide a substantial gain compared with the other simpler approaches.

HetNet scenario

In order to alleviate the spectrum scarcity problem, the future 5G networks will leverage on multi-connectivity supporting simultaneous connectivity across different technologies such as 5G, 4G, and Wi-Fi, multiple network layers, such as macro and small cells, and multiple RAT. This introduces a higher complexity in the management of the resources because the different layers and radio access technologies present different characteristics.

The multi-tenant ability introduced above needs to be extended in order to work even in this new scenario. Now we have to consider more possible allocation solutions because we share not only the spectrum resources but also the different technologies. This obviously introduces more complexity but even more flexibility, i.e., our algorithm has to decide not only about rejecting and accepting a request but even which technologies (among the available ones) are the most suitable for the service that the tenant wants to deliver to its users in respect of the QoS requirements (we could assign to a tenant even different technologies simultaneously).

5. Conclusions

The first public deliverable of WP4 presents the status of the 5G NORMA control and data layer design after the second of three planned design iterations.

The presented end-to-end c/d-layer architecture puts a clear focus on the RAN part. It internalized the 5G NORMA novel functionalities of adaptive (de)composition and allocation of mobile network functions as well as their software-defined mobile network control. The third enabling innovation, namely joint optimization of mobile access and core network functions, is not in the focus of this work package.

Most prominently, the architecture implements two novel functionalities that are crucial for future 5G networks, namely RAN slicing and multi-connectivity. RAN slicing extends network slicing into the RAN, thereby rendering them true end-to-end slices. RAN slicing is key to provide mobile network multi-tenancy as well as multi-service- and context-aware adaptation of network functions, the two 5G NORMA innovative functionalities. Multi-connectivity is orthogonal functionality to RAN slicing, i.e. it can be arbitrarily combined with it. Multi-connectivity adds a further dimension to service- and context-specific customization of radio access, supporting multi-RAT and HetNet environments as well as providing improved per user robustness and capacity.

The third and last design iteration will tackle left-overs from the second design iteration, both with respect to common topics such as RAN slicing and multi-connectivity, where discussions are still ongoing, as well as individually per partner, where innovations have not yet been fully integrated into the joint c/d-layer architecture, i.e., where their impact to and implementation on it still needs to be determined. Second, the integration and harmonization with the WP5 c/d-layer architecture within WP3 context will likely lead to necessary adaptations of the presented c/d-layer to be taken into account in the last design iteration. Finally, WP4 will target conclusive evaluations of its innovations. For several of them, evaluation results have been presented already in this deliverable. This evaluation work will continue to exemplify the benefit of individual innovations and the benefit of the proposed 5G NORMA architecture as a whole.

6. Annex A: Description of Function Blocks

Figure 6-1 shows the LTE processing chain in downlink and uplink, respectively, from a data layer point of view [DDM+13]. A similar classification can be found in [IJOIN-D22]. For clarity, some processing steps that are typically executed jointly due to their tight interdependency have been grouped into a single block. As shown, processing can be classified into cell and user specific processing. The cell-specific part consists of an analogue and mixed signal part of cell processing, described in Section 6.1, and digital baseband processing handled in Section 6.2. The remaining physical layer (PHY, or Layer 1) processing is already user specific and is covered in Section 6.3, which concludes the function blocks related to the physical layer. Layer 2 function blocks related to MAC are described in Sections 6.4 and 6.5, RLC in Section 6.6, PDCP in Sections 6.7 and 6.8, and eMBMS in Section 6.9. As 5G NORMA's functional decomposition spans not only RAN functions but also CN functions, Section 6.10 covers these non-access stratum (NAS) data layer functions and Section 6.11 covers the network side SDN-enabled transport.

The next Sections 6.12–6.14 cover cell and user specific RRC and the radio scheduler, i.e., RRM. These are the only distributed control layer functions, i.e., control functions that do not apply the SDMC concept. Finally, Sections 6.15–6.23 cover all SDMC-enabled control functions, which then fall under either common or dedicated control depending on the RAN slicing option in use.

Where applicable, according to Table 3-2, descriptions include additional or modified processing introduced to support 5G NORMA innovations.

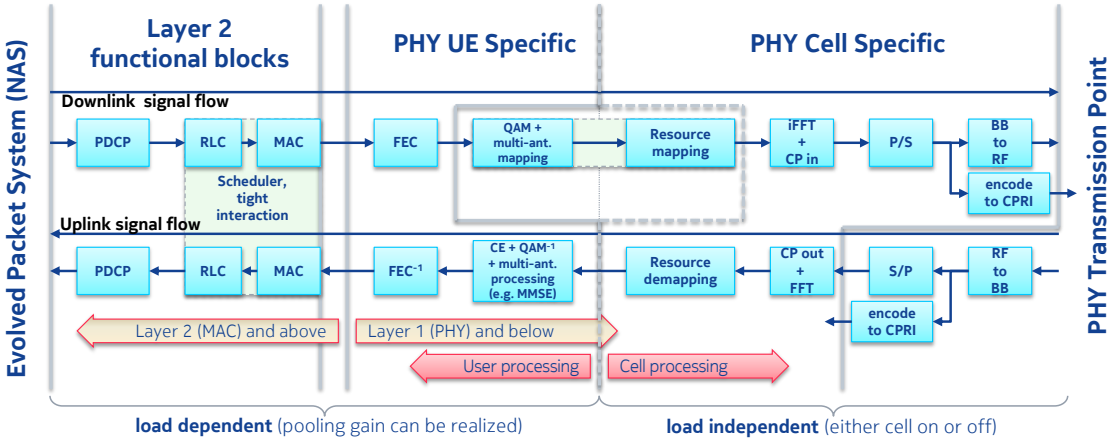


Figure 6-1: LTE Layer 1 and 2 processing chain and grouping of Layer 1 functions into PHY function blocks (dashed line if UE and cell specific PHY are co-located)

6.1. PHY Transmission Point

Description

The analogue and mixed signal processing for all signals transmitted and received via one transmission (and reception) point.

Details

More concretely, processing steps in downlink include D/A conversion, RF up-conversion, power amplification incl. pre-distortion, filtering and finally (directive) electromagnetic radiation via antennas, and in uplink (directive) absorption of electromagnetic radiation by antennas, filtering, low noise amplification, RF down-conversion and A/D conversion. For macro cells, mixed signal processing is typically separated from antennas into remote radio heads (RRH). EUTRAN (LTE) base station requirements are specified in [36.104] and signal generation in [36.211].

PHY Transmission Point is agnostic to layers, covering *both control and data layer*. Processing is synchronous, independent of channel coherence time (fixed processing chain latency), i.e. computational load is fixed for a given sampling rate, timing is fixed, which is a short constant processing time between antenna and *PHY Cell Specific*. Net transport bandwidth on the interface to *PHY Cell Specific* scales proportional to the number of active antenna ports and to the IQ sample resolution and sampling rate. Notably, the required bandwidth is user traffic load independent, i.e. it is the same for an idle cell and a fully loaded cell. *PHY Transmission point* is mostly *RAT agnostic*, so multiplexing of different RATs is supported, provided that these RATs are sufficiently compatible with respect to sampling rate, Tx power, Rx sensitivity, and filter requirements. Functional placement is primarily on *bare metal non-virtualised hardware* for the contained atomic functions D/A, A/D, PA, LNA, analogue filter, and mixer. The number of active antenna ports, Tx power, sampling rate, IQ sample compression, and carrier bandwidth impact power consumption and air conditioning requirements.

(De-)multiplexing of multiple PHY Cell Specific may be done in time domain in the case of downlink, i.e., adding (decompressed) time domain IQ samples in digital domain. In the case of uplink, selected time domain IQ samples are forwarded to selected *PHY Cell Specific* blocks. To simplify (de-)multiplexing, the basic maximum sampling rate should be common to all RATs that are multiplexed, i.e. all sampling rates should be an integer fraction of the basic sampling rate. For bandwidth efficiency, (de)multiplexing may be separated into a dedicated block to be executed on *virtualised hardware* within the Edge Cloud, co-located with multiple multiplexed *PHY Cell specific*. Thereby, a single multiplexed IQ sample stream is transported between the edge cloud and antenna instead of multiple IQ sample streams with one stream for each *PHY Cell Specific*.

Due to centralisation and virtualisation, monetary benefits arise through shared sites, space, housing, and processing for antennas, analogue and mixed signal parts as well as for the transport link, but possibly of higher capacity. Processing power requirements may be reduced because multiple dedicated processing chains are replaced by a single shared mixed signal processing chain. Some small additional processing is needed for (de-)multiplexing in the digital domain.

Interfaces

- **PHY Cell Specific.** Mandatory interface with an 1:n relation between function blocks *PHY Transmission Point* and *PHY Cell Specific*. Time domain (baseband) IQ samples are exchanged as synchronous fixed rate serial bit stream, which is typically transported over a special purpose interface such as CPRI [CPRI]. For efficiency, IQ samples may be compressed [GCTS12]. Future 5G RAT may support a packet based Ethernet/IP transport in case they can adapt to an increased latency introduced through receive/transmit buffers that convert the continuous time domain IQ samples stream from/to the discontinuous packet stream.
- **RRC Cell Specific.** Mandatory interface with 1:1 mapping of *RRC Cell Specific* to *PHY Transmission Point* for the exchange of control information.

Orchestrator input/output for function composition

- Setup parameters
    - antenna port on/off
    - sampling rate
    - sample resolution
    - compression scheme and parameters
    - carrier (center) frequency
    - bandwidth (tuneable analogue filters)
    - Transmission power
    - *PHY Cell Specific* instance to interface with

Controller input/output for function management

- Runtime parameters
    - cf. orchestrator input/output

6.2. PHY Cell

Description

(De-)multiplexing of *PHY UE Specific* and baseband signal generation including common PHY signals for one RAT or slice.

Details

In downlink, multiplexing *PHY UE Specific* in OFDM based systems (LTE and beyond) consists of mapping to resource elements (RE) and baseband signal generation (iFFT, CP insertion, and P/S conversion). If *PHY Cell Specific* and *PHY UE Specific* are not co-located within the same (edge) cloud, for efficiency reasons also modulation mapper, layer mapper and precoding from *PHY UE specific* should be executed in this block. Additionally, common signals for synchronisation, identification of cell, TP, or antenna (PSS, SSS, PBCH), and channel measurements are generated (CSI-RS, CRS, PRS). In uplink, demultiplexing consists of S/P conversion, FFT including CP removal and RE demapping. Signal generation and mapping for EUTRAN (LTE) are specified in [36.211].

PHY Cell Specific processing is synchronous, independent of channel coherence time (upper bounded processing pipeline latency of typically one TTI duration), and it handles *both control and data layer*. Processing is *RAT specific* within one *PHY Cell Specific* function block per cell (per each RAT and/or service). Functional placement is typically on *bare metal* employing specialised hardware for efficiency.

Computational load is fixed for load-independent signals and channels such as PSS, SSS, PBCH, and CRS. By contrast, computational load follows radio resource usage (user traffic load) in the case of (de)multiplexing uplink and downlink shared control and data channels (for LTE: PDSCH, PDCCH, PUSCH and PUCCH). The transport capacity needed for *PHY UE Specific* is independent of the number of antenna ports and follows the actual load and resource usage of a UE and is therefore much lower than that of *PHY Transmission Point* [DDM+13].

Different types of RE (de)mapping are used depending on the respective *PHY User Specific* block. For LTE, beside the already mentioned PDSCH, PDCCH, PUSCH and PUCCH, these are in downlink PMCH (for MBMS) and PDCH (for MIB transmission, further SIB broadcast as well as paging are already handled by PDCCH and PDSCH). In uplink, this is PRACH (RACHmsg1 detection) and for 5G small (sporadic) packet access SPTPmsg1/preamble [F5G-D31, F5G-D41].

5G small packet access in uplink assumes open-loop synchronisation like smart autonomous timing advance (ATA) based on downlink synchronisation signals [SW14]. For larger cells, where residual timing misalignment may exceed the CP length (CP-OFDM) respectively filter length (UF-OFDM), this may be handled by multiple overlapping FFTs depending on the actual cell size respectively the possible maximum misalignment [F5G-D31].

With respect to gain and cost for virtualisation and centralisation, monetary as well as processing benefits may arise primarily from sharing radio carriers, thereby processing load-independent signalling only once.

Interfaces

- **PHY Transmission Point.** Mandatory interface with n:1 mapping of *PHY Transmission Point* to *PHY Cell Specific* for supporting MIMO and CoMP JT.
- **PHY UE Specific.** Mandatory interface with 1:n mapping of *PHY Cell Specific* to *PHY UE Specific*. Exchange of frequency domain IQ samples. If cell PHY and UE PHY are not co-located, for efficiency reasons in DL instead exchange of FEC encoded (hard) user data bits and associated control for precoding, layer and modulation mapping.
- **RRC Cell Specific.** Mandatory interface with 1:1 mapping of *RRC Cell Specific* to *PHY Cell Specific*. Exchange of control information.

Orchestrator input/output for function composition

- Setup Parameters
    - For LTE cf. RRC parameters related to physical channels and signals [36.331].

Controller input/output for function management

- Runtime parameters
    - From the aforementioned setup parameters those parameters that are allowed to change during operation.

6.3. PHY User

Description

The generation of the baseband signal (in frequency domain for OFDM-based systems) from user data (DL) respectively decoding of baseband signals into user data (UL).

Details

In uplink, processing steps are channel equalisation, layer de-mapping, multi-antenna processing (e.g. MRC, MMSE or SIC receiver), modulation de-mapping, descrambling, control and data de-interleaving and demultiplexing, FEC decoding incl. HARQ combining (data only), transport block reassembly and CRC. In downlink, steps are CRC attachment, FEC encoding incl. possible segmentation into FEC code blocks and rate matching incl. bit selection (in the case of data according to HARQ redundancy version). If co-located with *PHY Cell Specific*, also scrambling, modulation mapper, layer mapper (i.e. multi-antenna mapper), precoding and DM-RS insertion. For LTE, multiplexing of control signalling and user data is done in *PHY Cell Specific*. For 5G user-specific in-resource control [PBF+16], multiplexing of control signalling and user data (of the single UE) is done here, while other shared 5G control signalling is multiplexed as in LTE by *PHY Cell Specific*. EUTRAN (LTE) multiplexing and channel coding are specified in [36.212] and PHY procedures in [36.213].

PHY UE Specific covers *both control and data layer*. It typically operates *synchronous*, once per TTI and processing time is upper bounded by one TTI duration. In 5G uplink, if *PHY Cell Specific* does multiple overlapping FFT per TTI (large cells), *PHY UE Specific* may operate multiple times per TTI and asynchronously, if the UE is currently in RRC Connected substate that lacks UL time synchronisation (UCA Enabled [ABAL16], RRC Connected Inactive [SMS+16], RRC Extant [F5G-D41]).

Processing is *specific to RAT and UE* configuration (RRC). For multi-connectivity, there may be more than one instance per UE per each RAT (in the case that the same UE connects to several RATs in parallel). The processing is stateless in DL and maintains a state in UL, where it employs a softbit buffer for HARQ combining. Conceptually it maintains state (in both UL and DL), holding RRC configuration, C-RNTI, UL power control state, current HARQ RV and further PHY related UE state. Since this and further UE states are needed by other function blocks such as *MAC Scheduling* and *RRC User*, too, it may be separated into a dedicated function block (database) holding all UE related configuration and control state.

Functional placement is typically on *bare metal* for efficiency reasons (channel equalisation, FEC decoding). Processing and transport latency towards UE depend on channel coherence time (link adaptation, HARQ RTT).

Gain of virtualisation/centralisation primarily stems from pooling gains, i.e. from exploiting statistical multiplexing among antenna sites, services and tenants, lowering the processing capacity that has to be available to cover peak loads and which directly translates into monetary (CAPEX) savings. Processing requirements can be influenced by link adaptation and scheduling. A more robust coding (lower modulation order, higher FEC rate) allows for successfully decoding with

less complex receivers (MMSE instead of SIC) and fewer turbo decoding iterations. This is especially useful to attenuate processing requirements during correlated load peaks and further reduces processing capacity requirements.

Interfaces

- **PHY Cell Specific.** Mandatory interface with n:1 mapping of *PHY Cell Specific* to *PHY UE Specific* for MIMO, CoMP JT and dual or multi-connectivity cell-less approach;
- **MAC UE.** Mandatory interface with 1:1 mapping of *PHY UE Specific* to MAC. Exchange of PHY transport blocks (in LTE a PHY TB equals a MAC PDU), including both control signalling and user data.
- **MAC Scheduling.** Mandatory interface with 1:1 mapping of *PHY UE Specific* to *MAC Scheduling*. Exchange of control information.
- **RRC UE Specific.** Mandatory interface with 1:1 mapping of *RRC UE Specific* to *PHY UE Specific*. Exchange of control information.

Orchestrator input/output for function composition

- Setup Parameters
    - For LTE cf. RRC parameters related to transport channels [36.331].

Controller input/output for function management

- Runtime parameters
    - From the aforementioned setup parameters those parameters that are allowed to change during operation.

6.4. MAC

Description

Data layer function block, which provides functionalities such as HARQ, AMC (allow to adapt the modulation and coding to the channel quality) and DRX (allow to improve UE battery life and energy saving).

Details

The MAC function block communicates with the physical layer in order to obtain info about channel quality and retransmission, with RLC layer regarding functionalities for HARQ; with RRC layer that is the responsible to enable the DRX functionality. In addition, it provides information to the UE regarding TX power, Network Ids, paging, and multiplexing in order to avoid collision or interference with the other UEs.

UEs in UCA enabled mode (substate of RRC_CONNECTED state) should be addressed within the UCA with one common Network Id, which is used by all access points belonging to the UCA. It has to be analysed whether a UE is provided with an additional identifier, which is used to address the UE if broadband data have to be transmitted. This additional identifier might be cell specific, i.e. comparable with the C-RNTI from LTE.

The MAC function block consists mainly of synchronous functions. However, there are some exceptions of functions which are asynchronous, such as Power Control and Network Identity management. Furthermore, it mainly refers to a “per user” functionality, i.e. one MAC instance operating per user, except for signalling of scheduling information. The MAC block is a RAT agnostic functionality, yet difficult to separate from other parts.

With the current architecture, the HARQ process requires that all uplink processing must be finished within 3 ms after receiving a subframe, so all uplink operations has to be performed in the same location. To relax this constraint, in 5G NORMA, we should develop a new approach that allows us to perform the uplink operations in a central processor (with more computational

power). MCS selection and scheduling also impose stringent timing requirements, which may be mitigated with predictive algorithms.

Interfaces

The MAC block interacts with the following blocks.

- **MAC Carrier Aggregation.** The exchanged information involves UE power control information as well as uplink coordination information.
- **RLC.** RLC PDUs are exchanged between MAC and RLC.
- **PHY User.** Signalling messages as well as MAC PDUs are transferred via this interface.
- **MAC Scheduling.** This interface involves scheduling information for each user.
- **RRC User.** Such interface involves exchange of radio bearer information.
- Self-organizing Networks. Interface with 1:1 mapping.

6.5. MAC Carrier Aggregation

Description

Data layer block, which coordinates the exchange of scheduling information as well as feedback information corresponding to the aggregated legs.

Details

Based on the exchanged scheduling information between the aggregated legs, a common priority handling is enabled by means of dynamic scheduling between UEs, as well as between logical channels of one UE. In addition, the *MAC carrier aggregation* (MAC CA) block carries out the 1x uplink functionality. This involves the utilisation of PHY control information which is associated with multiple channels in the downlink yet it is transferred in only one channel in the uplink. Such control information includes HARQ ACK/NACKs, scheduling requests, as well as channel quality indicator (CQI) reports regarding all the aggregated legs involved.

The relevant state of the art is summarised in [36.808], where the carrier aggregation operation is described in the LTE standards. Within the context of 5G NORMA, however, additional features will be accommodated. Such features are associated with the novel radio interfaces involved in 5G, hence their combined use into an inter-radio interface carrier aggregation scheme is not straightforward. The MAC CA block is thus specific to the employed RAT.

The MAC CA block is synchronous. In particular, it involves delay sensitive functionalities, such as the coordination of the 1x uplink feature, which renders it necessary that it runs close to the access. This implies that either the MAC CA block has to be co-sited with the PHY layer blocks, or the PHY layer blocks are interconnected with MAC CA via an ideal fronthaul link.

Interfaces

The MAC carrier aggregation block interacts closely with other blocks in the MAC layer. In particular, with

- **MAC Scheduling**. Scheduling and priority handling information is exchanged between *MAC Scheduling* and *MAC CA*. Such information is used for coordinating the aggregated legs.
- **MAC.** The exchanged information involves UE power control information as well as uplink coordination information. In addition, HARQ information is exchanged between *MAC* and *MAC CA*.
- **RLC.**

6.6. RLC

6.6.1. RLC Transparent Mode

Description

Data layer function block, which is dedicated to forwarding RRC information to/from lower layers. The term "transparent" stems thus from the fact that the RLC layer is not involved in RRC messages.

Details

RLC transparent mode (RLC TM) is an asynchronous block, i.e., not synchronous to the TTI level, which is utilised once per bearer, i.e., different bearers are associated to different RLC TM blocks. It is further dependent on the RAT used. Being transparent to the RLC layer, this block does not introduce noteworthy additional requirements.

Interfaces

- **RRC User, RRC Cell.** This interface is used to transfer RRC messages from *RRC User* block to the MAC layer and vice versa, via the RLC layer. Such RRC messages include paging messages as well as broadcast system information.
- **MAC UE.** Similar as above, it corresponds to the interface between RLC and MAC conveying RRC messages.
- **RRC mmW.** Special interface dedicated to transfer mmW-related RRC messages to/from RLC and RRC.

6.6.2. RLC Acknowledged/Unacknowledged Mode

Description

RLC acknowledged mode (AM) and unacknowledged mode (UM) primarily provide (re)concatenation and (re)segmentation of PDCP PDUs to adapt user data size to the amount of resources available to a radio bearer in a TTI.

Details

The RLC AM/UM block carries out the following operations: a) Reordering of the RLC PDUs which are out of sequence due to the retransmissions performed by the HARQ in the MAC layer; b) Coordination of duplicate RLC PDUs caused by misinterpretation of HARQ ACKs due to reception failure, so as to ensure that the PDCP PDUs are received only once; c) reassembling of the RLC PDUs for reconstructing the PDCP PDUs; d) Segmentation and concatenation of PDCP PDUs into appropriately selected sizes, based on information exchange with MAC and PDCP. In particular, the size of the RLC segments is determined dynamically, based on two factors: a) The RF transmission rate which depends on the instantaneous channel conditions; b) the volume of information accumulated in PDCP buffer.

- The main requirement of the RLC acknowledged/unacknowledged mode is that it has to be co-sited with the MAC layer, or interconnected with ideal backhaul. The explanation for such requirement is as follows: Although the interaction with PDCP is delay-tolerant, the interaction with MAC is delay sensitive. This stems from the time-varying nature of the wireless channel, which implies that the MAC scheduler is updated about the RF transmission rate in a frequent basis. This eventually leads to a large amount of control information exchange between RLC AM/UM and MAC. This implies that RLC AM/UM and MAC should be co-sited or interconnected with ideal backhaul; otherwise unnecessary overhead is created for the inter-site interface and the delay-sensitive interaction between RLC and MAC is violated. RLC AM/UM block is utilised once per bearer, and is specific to the employed RAT.

Interfaces

- **MAC UE**. HARQ information is exchanged between RLC AM/UM and MAC UE, leading to the potential reordering of RLC PDUs as well as to the coordination of duplicate messages. In addition, information related to the size of the data packets is also exchanged between MAC and RLC AM/UM.
- **MAC Scheduling**. Control information associated with scheduling is exchanged.
- **PDCP U/C**. This interface involves information which is used in conjunction with the information exchanged with MAC UE for the determination of packet sizes. Such information is related to the status of the data buffer in PDCP.
- **PDCP Split**. Similar as above, this interface is used for the case bearer split takes place.
- RRC User.
- **RRC mmW.** UE measurement and configuration of mm-wave cluster and UCA in downlink, UE measurement reception in uplink, 1:1 mapping

6.6.3. RLC Acknowledged Mode

Description

RLC acknowledged mode (AM) provides an additional re-transmission process, i.e., an automatic repeat request (ARQ) service, for increased reliability.

Details

The above mentioned retransmission process is known as outer ARQ (or simply ARQ), which operates on top of the MAC HARQ process. This results in additional robustness against missing packets. Moreover, if retransmission is indicated by the MAC layer, the RLC acknowledged mode block is associated with the re-segmentation of original data PDUs.

RLC acknowledged mode is asynchronous to the TTI level, and is specific to the RAT used. It is further utilised once per bearer. Moreover, similarly to the RLC AM/UM block, the RLC AM block has delay sensitive interaction with MAC. This incurs a timing requirement, implying that RLC AM/UM has to be co-sited with MAC or interconnected via ideal backhaul.

Interfaces

- **MAC UE.** The information exchanged between RLC AM and MAC involves retransmission messages, which append the HARQ retransmission process taking place in MAC with the outer ARQ retransmission process taking place in RLC.

6.7. PDCP

6.7.1. PDCP D

Description

The PDCP data layer function block is responsible to carry out all the functions like ROHC and data transfer procedures that are specific to the data layer.

Details

All the functions like Robust Header Compression (ROHC), ordered delivery and duplicate detection of the data packets are carried out with the PDCP-D block. ROHC compresses the header field of data packets. Compression is done using Least Significant Bit (LSB) encoding and Window based LSB encoding techniques. ROHC is configure to select different compression algorithms based on the higher layer protocols and their combination like IP, TCP/IP, UDP/IP etc. used. The operating mode of ROHC depends on the bearer type (DRB/ SLRB/ Split-Bearer), Uplink/Downlink and RLC modes (AM/UM). The ROHC channel is unidirectional with specific parameters defined for downlink channel and uplink channel. The mandatory parameters that are used for ROHC are Context Identification (CID), MAX_CID (Maximum value assign to CID configured by higher layers), LARGE_CIDS that is set depending on the value of MAX_CID,

PDU Type, SDU Type and PROFILES, that defines the type of data packet as TCP/IP or IP/UDP or No Compression etc. [36.323].

Side Link Radio Bearer (SLRB) corresponds to ProSe direct communication services; ROHC selects the compression algorithm corresponding to IP SDU's compression algorithm for SLRB's.

The function block is asynchronous to TTI and corresponds to the data layer category. The function block is instantiated once per RB. The functions and procedures are dependent on the type of radio bearer.

- The PDCP function block needs support for enhanced QoS based in service flow differentiation and service based radio bearer mapping. The parameters such as sequence number and hyper-frame number need to be exchanged for ordered delivery of packets. The value of discard timer can be set very high, therefore there are no strict latency requirements for the given function block.

Interfaces

- Interacts with RLC, PDCP C/D and PDCP split function blocks:
- **RLC.** ROHC selects the compression algorithm mode as unidirectional or bidirectional depending on the RLC operating modes.
- **PDCP C/D.** PDCP-D interacts with *PDCP C/D block* in order to carry out data procedures.
- **PDCP Split.** The interface is available only if the multi-connectivity supported UE is available and a split bearer is established.

6.7.2. PDCP C/D-Layer

Description

The functionalities and procedures corresponding to both control and data layer are included in the PDCP-C/D function block. It provides functions such as data transfer, sequence number maintenance, (de-)ciphering, integrity protection and verification.

Details

- The functionalities such as data transfer, sequence number maintenance and discard timer operate on both, data layer as well as on control layer. Different data transfer procedures are activated depending on bearer type (DRB, SRB, SLRB, Split-Bearer) and RLC modes.

The integrity protection and verification functionality operates only on c-layer with some exceptions defined by the device type, e.g., if it is a relay node, d-layer PDCP data is also integrity protected. The important parameters for the integrity protection and verification are DIRECTION that specifies the direction of transmission, COUNT defined by combination of PDCP SN (Sequence Number) and HFN (hyper Frame Number), a radio bearer identifier as BEARER, and integrity protection key provided by RRC layer as KEY (K_{RRCint}) [36.323]. Message Authentication Code for Integrity (MAC-I) is generated by integrity protection. MAC-I is compared to X-MAC generated in integrity verification using the Integration Protection Key.

Ciphering is performed on c-layer PDUs and data part of d-layer PDUs. In the case of relay nodes, ciphering is also applied to MAC-I of d-layer if it is integrity protected. Ciphering and integrity protection are closely related as the key for ciphering and integrity protection is generated by the upper layers, and provided by a single RRC message. The parameters for ciphering are similar to integrity protection and verification except that the KEY parameter consists of encryption keys.

Also, SL-(De)ciphering function is initiated by the ProSe function if SLRB bearer is used. The keys, ProSe Encryption Key (PEK), ProSe Traffic Key (PTK), ProSe Group Key (PGK) are generated with ProSe management function, and they are used as input parameters for ciphering algorithm.

- It is an asynchronous function block with respect to TTI and needs to be activated once per radio bearer. The main requirement of this function block is separation of d-layer and c-layer as most of the functions are only required for the c-layer. It also requires anchoring functions for flexible, on-demand data layer enhancements, including security and mobility on demand. The functions should manage inter-RAT packet routing by maintaining sequence numbers in the case of split bearer. Since PDCP functions operate asynchronously and the maximum value of discard timer can be set to infinity, there are no critical timing requirements on the functions [IJOIN-D31].

Interfaces

- **RRC User/Cell**. PDCP C/D block interacts with RRC User and RRC cell, in order to create, terminate, and re-establish PDCP context for radio bearers. Depending on the RB, different parameters and the security keys are transferred.
- **PDCP Split, PDCP D.** It also has an interface with *PDCP-D block* and conditional interface with *PDCP split* only in case of split bearer. Data procedures are carried out by such interfaces.
- **RLC.** It has bidirectional interface with *RLC*, as the RLC uses the same parameter context. The interface allows transmission of control and data packets, indication of pending packets, and acknowledgement.

6.8. PDCP Split

Description

PDCP Split Bearer block executes the functionalities of routing, reordering, and reordering timer operates in d-layer.

Details

The data packets need to be routed across the two simultaneously connected eNodeBs in LTE-A. Depending upon the buffer size the reordering window size is set to store the data packets. The reordering timer, t-reordering, is set to provide ordered delivery to split radio bearer and to avoid unnecessary NACKs that may be caused due to delayed delivery of data packets. The function also avoids loss of packets or duplication of packets by reordering the data packets that are stored in the reordering window. Once the reordering time is over, i.e. the parameter t-reordering is 0, all packets in the reordering window are delivered to the higher layers. The mandatory parameters for the function block are PDCP SN (Sequence Number), COUNT, REORDERING_WINDOW, Received Hyper Frame Number (RX_HFN), t-Reordering timer and Maximum Sequence Number MAX_SN.

The PDCP-Split function activates only in case of the split radio bearer. It also provides support for Inter RAT reordering of packets. The function block is asynchronous with respect to TTI, and is categorised under data layer. The function is instantiated once per split bearer. Common PDCP layer and presence of Split Radio Bearer is necessary for the activation of PDCP Split function block. Functional placement (based on requirements and interdependences) is typically in the edge cloud.

Interfaces

- **PDCP D** It operates closely with PDCP U block to carry out routing and reordering of data packets. Information such as frame synchronisation number, packet sequence number and reordering timer is transferred over the interface.
- **PDCP C/D.** The interface is active for ciphering of the data packets in case of split bearer.
- **RRC User.** It also interacts with RRC User (inter- RAT Link selection) function block if the UE is connected to more than one RAT.
- **RLC**. This interface enable transmission of data from PDCP to the lower layers in case of multi-connectivity.

6.9. eMBMS User

Description

Data layer function block, which performs the transmission of MBMS application data using the IP multicast address with the addition of SYNC protocol to guarantee that radio interface transmissions stay synchronised. Multimedia Broadcast Multicast Services (MBMS) offer support for broadcast and multicast services allowing the transmission of multimedia content (text, pictures, audio and video) utilizing the available bandwidth intelligently [23.246].

Details

eMBMS User is asynchronous respect to the TTI and handles *only data layer*. SYNC protocol depends on the maximum transmission delay from the Broadcast Multicast Service Center and the farthermost receiving node, the length of the synchronization sequence used for timestamping and the processing delay. Processing is *RAT specific*, depending also on the service deployment, with one eMBMS-U block per MBSFN (Multimedia Broadcast Single Frequency network) area. The eMBMS-U can be virtualized and centralized and the MBMS application data are forwarded to each TPs of the MBSFN area. Functional placement could be *Network Cloud* for general contents (e.g. national news service) or *Edge Cloud* for locally generated content (e.g. regional/city news service).

Computational load depends on maximum aggregate peak service rate (limited by the available resources at the RAN level). The Back/Front-haul bandwidth can vary according to the aggregate peak service rate.

- **Interfaces**
  - **NAS UE Data Layer**. Mandatory interface with 1:n mapping of eMBMS-U to NAS UE Data Layer for supporting the transfer of application data (GTPv1-U over UDP/IP).
  - **SON, self configuration**. (1:n)
- **Orchestrator input/output for function composition**
  - Setup parameters
    - Source IP address of MBMS Data
    - IP Multicast Service Groups
    - Maximum transmission delay in the MBSFN area (SYNC protocol)
- **Controller input/output for function management**
  - Runtime parameters
    - From the aforementioned setup parameters those parameters that are allowed to change during operation.

6.10. NAS

The NAS function block subsumes all data layer functions of the non-access stratum. In 4G these functions are provided by the S-GW and P-GW network entities. NAS related data layer functions are not in focus of WP4. The NAS block has been added to show the interfacing of RAN network functions towards the NAS.

6.11. Transport (SDN)

Description

Connectionless routing within RAN based on SDN configuration.

Within an UCA a connectionless routing of downlink small data packet from the anchor node to the best serving node and vice versa in uplink has to be established for each UE.

Details

For each UE a dedicated UCA is defined by the SON control which is communicated to RRC User/RRC mmW. The SON control will also inform the RAN Paging about the access nodes belonging to the UCA. The RAN Paging will trigger the SDM Controller (SDM-C) to set up:

- dedicated data paths between all nodes of the UCA and the anchor node for uplink data transmission,
- and one dedicated data path from the anchor node to one (best server) node of the UCA.

In case of a movement of an UE towards a better serving node inside the UCA, the RRC User/RRC mmW - informed by the UE – will inform the SON control about the new anchor node, which will trigger the RAN Paging to update dedicated data paths.

In case the UE selects a new node outside of the existing UCA, the SON control – triggered by the RRC User/RRC mmW – will define a new UCA for this UE and will also trigger data paths updates by towards the RAN Paging.

These data paths will be used on demand to transfer small data packets in uplink and downlink inside the UCA. A low computational load is expected for setup of connectionless transmission paths in downlink and uplink. A fast modification of the downlink transmission path in case of UE mobility, i.e. a received indication of new best serving node, is required. The procedure is asynchronous, as it is trigged by the UE which signals a new best serving node.

Orchestrator input/output for function composition

None.

Controller input/output for function management

- Interfaces between the RRC User/RRC mmW and the SON controller and between the SON controller and RAN Paging
- Interface from RAN Paging to towards SDM-C controller to trigger a setup of dedicated data paths.
- Output from SDM-C controller to all Transport (SDN) inside the UCA: setup and update of connectionless transmission paths, further details need to be evaluated.

6.12. RRC Cell

Description

The cell specific RRC refers to a control layer function block, which handles control layer signalling protocols associated with broadcasting system information, including NAS common information and information relevant to UEs in RRC_IDLE, e.g. cell (re-)selection parameters, neighbouring cell information and information (also) applicable for UEs in RRC_CONNECTED, e.g., common channel configuration information.

Details

UE acquires system information mainly containing *cell access-related information* to use transmission channels in both RRC_CONNECTED and RRC_IDLE modes. The *Cell Specific RRC* protocol controls the UE behaviour by transferring common NAS-related information, i.e. NAS information which is applicable to all UEs, and AS-related information per cell [36.331].

The main function is the broadcast of system information including NAS common information applicable for UEs in RRC_IDLE and RRC_CONNECTED. The information is mapped directly to the logical Broadcast Control Channel (BCCH). Parameters include in the Master Information Block (MIB), the DL system bandwidth and system frame number (SFN); in the SIB1 the cell access-related information, e.g. tracking area code and cell identity, cell-selection information, e.g. minimum required Rx level in the cell, scheduling information, TDD configuration. There are further SIBx defined up to SIB11 including information on common radio resource configuration information, common information for intra/inter-frequency and inter-RAT cell re-selection.

Interfaces

- **PHY TP.** Mandatory interface with 1:1 mapping to exchange of control information,
- **PHY Cell Specific.** Mandatory interface with 1:1 mapping to carry the master information block (mapped on BCCH), which consists of limited number of most frequently transmitted parameters essential for initial access to the cell, on the Physical Broadcast Channel (PBCH),
- **MAC Scheduling.** Mandatory interface with 1:1 mapping to exchange multiplexing information together with unicast data,
- **RLC transparent mode.** 1:1 mapping of system information as delay-sensitive applications with no reliability concerns onto BCCH,
- **NAS Event Control Layer.** Mandatory interface to exchange the relevant c-layer information between cells and core network; n:1 mapping
- **eMBMS.** Mandatory interface with n:1 mapping of RRC Cell to eMBMS-C for supporting the configuration of SIB13, MCHs scheduling and service information (GTPv1-U over UDP/IP).

Orchestrator input/output for function composition

Logical address of the function (physical address(es) of hosting entity(ies)), computational requirements of the function, link requirements for signalling. This will be refined based on further discussion in WP5.

Controller input/output for function management

- The information regarding each the radio resources for each slide and their requirements are provided to RRC from the controllers, i.e., SDM-C and SDM-X. It is suggested that as an alternative SDM-C can host the centralized radio resource management (explained in second part) algorithm. Then, SDM-C will also provide policies regarding cell configurations (e.g., TDD-paterns).
- Exchanging information on the RRC signalling integrity protection and RRC signalling ciphering for mobility;
- Acquiring the UE point of attachment via RRC signalling and store the UE's point of attachment to virtual AAA.

6.13. RRC User

6.13.1. RRC User

Description

This is associated with a control layer function block, which handles the UE management and control. The RRC User entity also takes care of multiplexing data packets coming from the upper layers into the appropriate radio bearer.

Details

- *RRC User Specific* controls the UE behaviour in RRC-IDLE and RRC_CONNECTED by transferring RRC messages, dedicated NAS information (RRC_CONNECTED), and paging (RRC_IDLE). The functionality includes RRC connection control, which covers all procedures related to the establishment, modification, and release of an RRC connection; inter-RAT mobility including the transfer of RRC context information; and measurement configuration and reporting. The *QoS innovation* will provide a new QPS (QoS Parameter Set) and new dynamic bearer which need to be controlled accordingly by *RRC User Specific*.

Interfaces

- **MAC Scheduling.** 1:1 mapping for exchanging multiplexing information.
- **MAC.** 1:1 mapping to exchange UE-relevant c-layer information like DRX

- **RLC Transparent Mode.** 1:1 mapping of delay-sensitive user specific applications with no reliability concerns
    - Paging mapped onto PCCH
    - Dedicated RRC messages using CCCH applied for RRC connection establishment
        - SRB0
- **PDCP U/C.** Mandatory interface with 1:1 mapping to exchange c-layer information to be integrity-protected and ciphered
- **NAS UE-Specific Control and Data Layer.** n:1 mapping to exchange c/d-layer information of UEs with core network
- **NAS Event-Control Layer.** n:1 mapping to exchange c-layer information of UEs with core network
- **Inter RAT/Link Selection.** QoS requirements, latency requirements and UE preferences are exchanged over this interface.

6.13.2. RRC for mm-Wave

Description

Additional functionalities for user-centric RRC which are not covered by the RRC User description, i.e. the following details are related to:

- mm-wave transmission points controlled by 5G coverage cell and
- UCA for short data packet transmission

Details

For each UE, a definition of a cluster of mm-wave access points is required. The cluster will be defined by a SON function inside e.g. the access cloud. The RRC for mm-wave will support the SON function by provisioning of UE measurements of surrounding mm-wave access points. After receiving the cluster definition *RRC for mm-wave* has to configure the UE which includes the number of mm-wave access points and detailed further info, e.g. information about mm-wave beams measurement configuration (number of beams, timing, power etc.) towards UE. During mobility of the UE, the RRC for mm-wave has to take the decision which mm-wave access point should serve the UE and has to trigger the PDCP Data forwarding to one or several mm-wave access points. With respect to timing requirement, a fast reaction on UE measurements, e.g. selection of new serving mm-wave AP is required. In addition, a high speed X2 backhaul for forwarding of UE data to the mm-wave APs should be provided.

For each UE, a definition of an UCA is required. The UCA will be defined by a SON function inside e.g. the access cloud also based on UE measurements. The *RRC for UCA* has to configure the UE with the UCA definition. Information about the UE context has to be provided to all other RRC for UCA instances inside the UCA. In addition, a UCA-specific paging is required in case download data arrives and the position of a UE inside a UCA is not known, e.g. due to not received best server indication of the UE. Inside the UCA, a connectionless routing of data is proposed, which has to be set up by the SDM Controller inside (SDMC) towards Transport (SDN).

All above mentioned features (for mm-wave and UCA) are belonging to the c-layer and d-layer, have to be implemented for each UE, i.e. UE centric, and are asynchronous, as they are based on UE measurements.

Interfaces

- **RLC AM/UM, RLC TM.** UE measurement and configuration of mm-wave cluster and UCA in downlink, UE measurement reception in uplink, 1:1 mapping.
- **RLC TM.** Transfer of mmW-related RRC messages to/from RLC and RRC.
- **PDCP U/C.** UE measurement and configuration of mm-wave cluster and UCA in downlink, UE measurement reception in uplink, 1:1 mapping.

Transfer of mmW-related RRC messages to/from RLC and RRC

Triggering of PDCP forwarding to mm-wave access points: internal in case of RRC for mm-wave is implemented in cloud or "X2" interface between dedicated nodes, 1:1 or 1:n mapping depending on configuration

- **RRC mmW.** Distribution of UE context: internal in case of RRC for UCA is implemented in cloud or "X2" interface between dedicated nodes. 1:n mapping seen from the transmitting node, 1:1 mapping seen from the receiving node.
- **MAC UE.** Exchange of radio bearer information.

Orchestrator input/output for function composition

- none.

Controller input/output for function management

- **RAN paging.** Triggering of paging in RAN: internal in case of RRC for UCA is implemented in cloud or "X2" interface between dedicated nodes, 1:n mapping seen from the transmitting node
- **SON.**
  a: UE measurements to trigger a UCA or mm-wave cluster definition to SON,
  b: Reception of defined UCA and mm-wave cluster from SON.
  c: UE movement to re-configure Transport SDN via RAN Paging and SDM-C

6.13.3. RAT/Link Selection

Description

The Inter RAT link selection block enables link selection and packet scheduling, if the UE is simultaneously connected to two or more RATs.

Details

The function block provides connection of UE to two or more RATs at the same time. The block enables inter-RAT link selection operation, i.e. the best link is selected across different RATs in order to achieve maximum throughput. It maps the service request to RATs considering radio signal strength (RSS), reference base-station efficiency, traffic, latency and security requirements. It acts as an anchor point to dynamically select multi-connectivity operational modes (data duplication and data diversity) in the case of split bearer, and hence also allowing packet scheduling across multiple RATs.

The function block belongs to both the d-layer as well as c-layer. It is asynchronous with respect to TTI and operates on top of the PDCP function blocks. The block is activated once per multi RAT connected UE. The block interacts with PDCP-C/D, PDCP Split bearer and RRC User function block. The requirement for the function block is presence of common PDCP layer across RATs and UE support for multi connectivity. Functional placement is typically in the edge cloud.

Interfaces

- **PDCP Split:** The function has south Bound Interface with PDCP- C/D, PDCP Split functional blocks. The information about the split bearer established between two RATs need to be provided to inter- RAT support function block so as to allow data transfer and service mapping. The parameters such as the list of established RB connections, available and connected RATs to UE are exchanged over the interface.

Orchestrator input/output for function composition

- To be determined in conjunction with WP5.

Controller input/output for function management

- Input from traffic monitoring if the change in the system load exceeds a predefined threshold.

6.14. MAC Scheduling (RRM)

Description

This block is responsible for scheduling the transfer of user data and control signalling in downlink and uplink subframes over the air interface.

Details

More specifically, this is a control and data layer function block, which uses information from upper (e.g. RLC and RRC) and lower (e.g. PHY) layers and provides the following services as specified in [36.321]:

- Mapping between logical channels and transport channels;
- Multiplexing/de-multiplexing of MAC SDUs belonging to one or different logical channels into/from transport blocks (TB) delivered to/from the physical layer on transport channels;
- Scheduling information reporting;
- Avoiding padding and segmentation;
- Mapping physical channels on logical channels;
- Priority handling between logical channels of one UE;
- Priority handling between UEs by means of dynamic scheduling;
- ICIC/CoMP;
- D2D support;
- BS power control.

In the downlink, the MAC Scheduler allocates resources which are used to send transport blocks to specific UEs via the DL-SCH transport channel. These downlink transport blocks contain:

- Upper layer data: All upper layer data, whether this is user data or signalling messages, will be fed into the MAC sublayer through an RLC logical channel. Whenever one of these logical channels has any data to send over the LTE air interface, a Buffer Status indication is provided by RLC to the MAC Scheduler
- MAC layer data: This can either take the form of a HARQ retransmission or MAC Control Element. In both cases, the MAC layer will provide an indication to the scheduler that it has data ready to send over the LTE air interface.

In the uplink, the MAC Scheduler allocates resources to specific UEs which enable them to set up and send transport blocks to the eNodeB via the UL-SCH transport channel. Each UE needs to request the uplink resources needed to send a transport block. This can be done in different ways:

- When a UE does not have an active connection in a cell, it performs the Random Access Procedure to join (or re-join) the cell.
- When the UE has uplink control resources on the PUCCH physical channel, it sends a Scheduling Request indication to request uplink resources.
- When the UE already has resources to send data in the uplink on the UL-SCH transport channel, it may indicate to the scheduler that it has more data to send, and therefore requires more uplink resources. The UE does this by sending a Buffer Status Report MAC Control Element.
- Based on HARQ feedback at the eNodeB, the scheduler might need to allocate uplink resources for HARQ retransmissions by the UE.

The output of the scheduling algorithm is a scheduling assignment per uplink and downlink subframe, which is signalled to UEs on PHICH (to schedule non-adaptive HARQ retransmissions or to suspend or terminate HARQ processes) or (E)PDCCH (to schedule initial HARQ transmissions or adaptive HARQ retransmissions incl. resuming suspended HARQ processes) physical channel.

- The downlink assignment information will be used by the eNodeB MAC layer to generate suitably sized transport blocks which are passed to the eNodeB PHY layer, and used by the PHY layer to FEC encode and map the transport blocks accordingly into PDSCH time

and frequency resources and spatial layers, before it is subsequently transmitted to the UE.

- The uplink assignment information indicates which UEs have been allocated uplink resources within the subframe and the MCS to apply. It is also used by the eNodeB PHY layer to identify uplink transmissions in the corresponding uplink subframe and forward it to the correct MAC instance.

Before any user data can be transferred over the air, the scheduler needs to be configured with a cell specific configuration, a UE specific configuration and finally a bearer specific configuration for each active UE bearer. Reconfiguration of each UE and bearer is also possible.

The PHY layer provides services such as data transfer, signalling of Scheduling Request and measurements, e.g. Channel Quality Indication (CQI). All this information is needed in order to tune properly scheduling functionalities. The channel mapping of logical channels to transport channels depends on the multiplexing that is configured by RRC.

MAC Scheduling is also responsible for scheduling the cell's radio resources used in the downlink and uplink while providing the required QoS for all active radio bearers.

The *MAC scheduling* block is delay sensitive. So in order to meet these constraints we need to decentralise the uplink processing placed in the same location of MAC or develop new algorithms in order to centralise the uplink processing, complying with the delay constraints. It seems difficult that this time constraint can be relaxed in any way, so this functionality should probably run close to the access – maybe the scheduling functionality could be split and part of it could run further from the access.

Providing a service flow with high data rate and ultra-reliability, the service flow may be split using a multi-connectivity setup where each multi-connectivity interface has an independent RLC. Furthermore, the QoS innovation will provide a new QPS (QoS Parameter Set) and a new dynamic bearer that needs new scheduling functions to provide suitable scheduling decisions.

Interfaces

- **PHY UE Specific.** Mandatory interface with 1:1 mapping. Exchange of signalling of Scheduling Request and measurements (e.g. CQI), MCS, HARQ RV, resource mapping; in UL CRC result, Rx power, interference level.
- **RRC Cell, RRC UE.** Mandatory interfaces with 1:1 mapping. Exchange of multiplexing information.
- **MAC Carrier Aggregation.** Scheduling and priority handling information.
- **MAC UE.** Provide scheduling information for each user, i.e., MAC PDU size, MAC CEs, in UL additionally info from received MAC CEs BSR and power headroom report.
- **RLC AM/UM.** Exchange of control information. RLC buffer status (input to scheduling decision), RLC PDU size (output); RLC AM.
- Self-organizing Networks. Interface with 1:1 mapping.
- **eMBMS.** Mandatory interface with 1:n mapping of eMBMS-C to MAC Scheduler for supporting the configuration of MTCHs Scheduling Information (SCTP over IP).
- **RAN paging.** Interface to RAN paging so as to trigger the RAN paging from HARQ process, i.e. if a data transmission inside the UCA towards a UE is not confirmed. 1:1 mapping for this interface, multiple instances for different UEs may run in parallel.
- **SDM-X.** Interface to SDM-X policies in order to exchange the information regarding resources sharing, e.g., the location and the amount of resources dedicated to each tenant in a determined time interval (input from SDM-X policies) and the number of users belonging to each tenant and their traffic demand (output to Multi-Tenancy Scheduling). Through this interface it can communicate with the Multi-Tenancy Scheduling that is the responsible for managing the resource sharing among multiple tenants.

Orchestrator input/output for function composition

- The real-time available spectrum resources and spectrum sharing should be sent to orchestrator for oversight coordination.

Controller input/output for function management

- For further investigation in WP5

6.15. Multi-tenancy Scheduling

Description

This block is responsible for coordinating resource sharing among multiple tenants.

Details

The Multi-tenant scheduling functionality is responsible to controls the underlying scheduling to allocate dynamically resources to different slices/tenants. A new tenant, which needs to provide a particular service in a certain area for a limited period, can send a request to this module in order to obtain the amount of resources needed to respect the service requirements. When received a new request, the Multi-tenant scheduling application, given the actual network load information, decide if reject or accept it. In the latter case, it controls through the SDM-X the MAC scheduling (RRM) in order to serve the tenant's users properly.

Interfaces

- **PHY UE Specific.** Mandatory interface with 1:1 mapping. Exchange of signalling of Scheduling Request and measurements (e.g. CQI)
- **RRC Cell, RRC UE.** Mandatory interfaces with 1:1 mapping. Exchange of multiplexing information.
- **MAC Scheduling (RRM).** Mandatory interface with 1:1 mapping. Exchange of Scheduler parameters (amount of resources per tenant).
- **SDM-X.** Exchange of information in order to control the MAC Scheduling (RRM) according to the resource sharing policy adopted.

Orchestrator input/output for function composition

- The real-time available spectrum resources and spectrum sharing should be sent to orchestrator for oversight coordination.

Controller input/output for function management

- For further investigation in WP5

6.16. mMTC RAN Congestion Control

Description

Grouping mMTC devices into context-based clusters and schedule their RAN procedures in subframes, to reduce the RAN congestion rate.

Details

Details are ffs and will be included in D4.2.

Interfaces

- **RRC User.** Details are ffs and will be included in D4.2.
- **RRC Cell.** Details are ffs and will be included in D4.2.

Orchestrator input/output for function composition

Details are ffs and will be included in D4.2.

Controller input/output for function management

Details are ffs and will be included in D4.2.

6.17. QoS Control Description

The QoS control function block is intended to be implemented as application over the SDM-C/X in the control layer. This function will be in charge of the network monitoring and configuration in real time through open interfaces that interact with the SDM-X, SDM-C and the control radio stack.

Details

The QoS control function block will provide real-time control of traffic flows in 5G NORMA based on the QoS parameters defined for a specific service, allowing the network to be reorganised over time according to service evolution requirements, network slicing configuration and demand load. Reconfiguration of the network resources is envisaged with the suitably named QoS management function block defined in the management layer.

Interfaces

This function will interact with the following function blocks:

- Control Layer: SDM-C/X, MAC scheduling. Used to receive monitoring data (events) from the network elements in the case of network monitoring, and to send configuration information in the case of network configuration.

Controller input/output for function management

QoS management function block manages a flexible initial set of parameters at SLA level that will be treated by the different orchestrators and translated to specific dynamic QoS parameters at network level managed by the QoS control function block.

6.18. Self-Organizing Networks

Description

Self-Organised-Network (SON) handles the network management operations mostly in a distributed manner. However, it can be applied via a centralised supervision (hybrid solution) or fully centralised fashion. The scope of SON functions covers different features: *i*) Self-configuration, *ii*) Self-Optimisation, *iii*) Self-Healing, and *iv)* User Centric Connection Area (UCA) and mm-wave cluster configuration. In the remainder of this section, we present the above features in detail. We also present a new functionality (c.f. part iv), which is appropriate for the UCA approach.

i) Self-configuration

Details

This function specifies the parameters for the initial configuration during the integration within an existing network. Next, we highlight the requirements of the self-configuration feature of self-organizing networks.

Inf-N interface must be placed amongst the OAM and different base stations. Those functions can be applied to heterogeneous networks (macro-cell eNodeB, mm-waves cells, small cells, etc.). There might be a gap to support evolving RAN technologies. SDN centralised c-layer can help towards this direction.

Support for network slicing and multi-tenancy. The multi-tenancy support over a large area may involve a new set of base stations, which need to be logically connected to the existing network domain and advertised to neighbouring base stations through SON functions. The self-configuration functions needs to instruct the neighbouring path and build the X2 interface in a dynamic way (short-term time window).

Support for cell-less dual-connectivity network. SON functions are exploited to dynamically create a dual-connectivity area (multi-service area), wherein UEs may take advantage from being

connected to multiple eNodeB (or small cells, or mm-waves access points) simultaneously. The multi-service area could be easily adjusted by monitoring and changing the transmission power of the eNodeBs.

Configuration has been handled in a centralised way but support for new technologies would require extensions that can be handled, e.g. by SDN

Interfaces

- **NAS Event-Control Layer.** (n:1) SON based control procedures
- **NAS Core Network.** (n:1) SON based control procedures
- **eMBMS-U.** (n:1) MBMS session management
- **RRC User.**

Orchestrator input/output for function composition

- For further evaluation in WP5

Controller input/output for function management

- For further evaluation in WP5

ii) Self-optimisation

Details

This set of functions addresses different features, such as optimisation of cellular coverage, handover operations (including mobility operations aiming at optimal load balancing), network capacity optimisation, energy saving, and interference mitigation/orchestration.

X2 interface must be in place among neighbouring base stations. In heterogeneous networks, an optional centralised control can assist this operation. eNodeBs can be switched off (no user associated and handover requests rejected), and waken-up when the traffic load increases. Cell coverage must be guaranteed by other cells. Coordination among SON functions is essential to make sure that there are no conflicting goals. There might be a gap to support evolving RAN technologies. SDN centralised c-layer can help towards this direction.

Network slicing support among different base stations. Load balancing operations may help the network slicing to easily improve the overall network performance by means of SON functions.

Support for cell-less dual-connectivity network. SON functions are exploited to dynamically create a dual-connectivity area (multi-service area), wherein UEs may take advantage from being connected to multiple eNodeB (or small cells, or mm-waves access points) simultaneously. The multi-service area could be easily adjusted by monitoring and changing the transmission power of the eNodeBs.

Interfaces

- **NAS Event-Control Layer.** (n:1) SON based control procedures
- **NAS Core Network.** (n:1) SON based control procedures
- **MAC Scheduling.** (1:1) Scheduling operations based on SON basic functions
- **MAC UE.** (1:1) MAC functions requiring SON operations
- **MAC Cell.** (1:1) MAC functions requiring SON operations

Orchestrator input/output for function composition

- For further evaluation in WP5

Controller input/output for function management

- For further evaluation in WP5

iii) Self-healing

Details

These functions help in detecting and recovering from network failures. For instance, coverage problems are dynamically addressed when network topology variations affect the network.

An optimised SON based allocation of access points belonging of a UCA will minimise the radio and core network signaling overhead related to connection management (idle/active transitions) and to mobility (paging, handover).

An optimised SON based configuration of mmWave access points improves the transmission quality with respect to blocking effects caused by sudden user movement or obstacles entering the transmission path.

An optimised SON based configuration of small cells (or mmWave access points) improves the multi-tenancy support over large area by exploiting the spatial domain.

Interfaces

- **NAS Event-Control Layer.** (n:1) SON based control procedures
- **NAS UE Specific Control Layer.** (n:1) SON based control procedures
- **NAS Core Network.** (n:1) SON based control procedures
- **MAC UE.** (1:1) MAC functions requiring SON operations

Orchestrator input/output for function composition

- For further evaluation in WP5

Controller input/output for function management

- For further evaluation in WP5
-

iv) UCA and mm-wave cluster configuration

Description

A new SON functionality is required for the definition of a UCA and an mm-wave cluster, which has to be individually defined for each UE.

Details

A User UCA and a mm-wave cluster consists of a set of cells selected by the 5G-RAN. It can be defined based on the neighbour cell measurements of the UE, but should not be limited only to that. A SON functionality taking into Neighbour Relation Table (NRT) or even anticipatory techniques like mobility tracks of UEs (e.g. movement within pedestrian area) will provide a better cluster definition. This SON functionality requires low computational load and has relaxed timing requirements. It is asynchronous as it is trigged ones if a UE is addressing the 5G system and by further UE mobility. Multiple instances have to setup as for each UE a UCA or mm-wave cluster has to be defined.

An optimised SON based allocation of access points belonging of a UCA will minimise the radio and core network signalling overhead related to connection management (idle/active transitions) and to mobility (paging, handover).

An optimised SON based configuration an mm-wave access points improves the transmission quality with respect to blocking effects caused by sudden user movement or obstacles entering the transmission path.

Interfaces

- **RRC for mm-wave and UCA.** UE measurements of surrounding cells as input, definition of UCA and mm-wave cluster as output, 1:1 mapping
- **RAN Paging:** Transmitting of an update request for connectionless data transmission inside a UCA to be forwarded via SDM-C to Transport (SDN)

Orchestrator input/output for function composition

- **none.**

Controller input/output for function management

- **none**

6.19. RAN Paging

Description

To reduce signalling messages on the air interface and towards the core network, 5G NORMA proposes a RAN paging approach, i.e. inside an UCA, in addition to a paging in a larger tracking area.

Details

UCA enables a UE based mobility inside the UE specific UCA. Each update of the best server is signalled by the UE based on one connectionless short data message. Compared to a data transmission based on a scheduling request and a scheduled uplink transmission such an update might not reach the addressed access point. In this case, the data transmission by the last known best server will fail and consequently a paging inside the UCA will be required. This is defined as RAN paging in contrast to the network based paging controlled by the MME, in which many more access points compared to the access points within a UCA are involved.
The RAN paging is also responsible to trigger via the SDM-C the setup of a connectionless data transfer inside the UCA, which is based on Transport (SDN).
The RAN paging is an asynchronous c-layer procedure as it is based on detection of missing UE allocation to best server access point and UE movements inside the UCA, it generates low computational load, it has a relaxed timing requirement, and it has to take into account the DRX cycle or UE to be paged.

Interfaces

- **RRC mmW (requesting node inside UCA).** Reception to triggering of a paging inside the UCA for a dedicated UE inside the coverage area of the requested access node, 1:1 for this interface, multiple instances for different UEs may run in parallel.
- **RRC mmW (nodes which have to initiate paging).** Command to carry out a UE paging by each node inside the UCA. 1:1 mapping for this interface, multiple instances for different UEs may run in parallel.
- **SON.** Reception of an update request for connectionless data transmission inside a UCA to be forwarded via SDM-C to Transport (SDN)

Orchestrator input/output for function composition

- To orchestrator in the case RAN paging is not successful and a paging within a larger area than the UCA will be required. Further details need to be analysed.

Controller input/output for function management

- None.

6.20. eMBMS Control

- **Description**
  Control layer function block which performs the admission control and allocation of the radio resources, UE counting procedure, MBMS session management (initiating the MBMS session start and stop procedures), allocation of an identity and the specification of QoS parameters associated with each MBMS session.
- **Details**
  eMBMS-C is asynchronous respect to the TTI and handles *only c-layer*. eMBMS-C depends on the ARP (Allocation and Retention priority, a QoS parameter) and on the result

of counting procedure. Processing is *RAT specific*, depending also on the service deployment, with one eMBMS-C block per MBSFN (Multimedia Broadcast Single Frequency network) area. Functional placement could be in *Network Cloud* and/or *Edge Cloud*.

Back/Front-haul bandwidth is only used to send signalling traffic. The eMBMS-C can be fully virtualized and centralized but it is possible having some instances of eMBMS-C also at the edge to optimize the signalling traffic distribution (e.g. edge cloud which is serving the nodes belonging to the same MBSFN).

- **Interface**
  - **RRC Cell**. Mandatory interface with 1:n mapping of eMBMS-C to RRC Cell for supporting the configuration of SIB13, MCHs scheduling and service information (GTPv1-U over UDP/IP).
  - **MAC Scheduler**. Mandatory interface with 1:n mapping of eMBMS-C to MAC Scheduler for supporting the configuration of MTCHs Scheduling Information (SCTP over IP).
  - **NAS Event-C**. Mandatory interface with 1:n mapping of eMBMS-C to NAS Event-C for supporting the MBMS session management, including QoS parameter (GTPv2-C, SCTP over IP).
- **Orchestrator input/output for function composition**
  - Setup parameters
    - Area configuration
    - Subframe allocation
    - Service Groups configuration and QoS
- **Controller input/output for function management**
  - Runtime parameters
    - From the aforementioned setup parameters those parameters that are allowed to change during operation.

6.21. NAS Control

6.21.1. NAS UE Specific and Data Layer

Description

NAS UE specific d-layer function block refers to the user perspective functions and procedures related to the d-layer which are triggered by the NAS UE-specific c-layer functions.

Details

- The block is composed of two parts: *i*) User-agnostic, which are out of the UE control, *ii*) User-centric functions, involving the communication between the UE and the CN. The user-agnostic functions include
  - Deep packet inspection: this function examines the data part and, possibly, the header of a packet and it is triggered by the P-GW/S-GW entity.
  - Lawful interception: this function physically makes a copy of (part of) the traffic and it is performed by a network operator / access provider / service provider through the P-GW/S-GW.
  - Mobility anchor management: this function provides related control and mobility support between GPRS core and the 3GPP anchor function of S-GW (by means of the S4 interface). These operations are triggered by the MME.
  - UL/DL charging: this function charges the ID communicated by the UE through the NAS control layer functions.
  - Dedicated Bearer instantiation: this function is triggered by the MME and executed through the P-GW when a dedicated bearer is required.

- Packet routing: this function performs the process of determining and using, in accordance with a set of rules, the route for transmission of a message within and between the PLMN(s). This is performed by the P-GW.
- Packet marking: this function sets the DiffServ Code Point (DSCP), based on the QCI of the associated EPS bearer. The DSCP is set by the endpoints, such as the eNodeB and the P-GW.
- Data forwarding in RAN: this function forwards the PDCP packets during hand-over and for dual connectivity from one radio node to another radio node using the X2 interface. It is performed by the eNodeB.

The user-centric functions include

- NAS-based IPv4 address allocation: this function assigns an IP after the default bearer allocation. It is performed by the MME through the P-GW.
- DHCP (DHCPv4 and DHCPv6): this function delivers the IP configuration information to the UE. It is performed by the P-GW after the default bearer allocation.

Interfaces

- **NAS UE Specific Control Layer.** (n:1).
- **NAS Event-Control Layer.** (1:1).
- **RRC User**. (n:1).
- **eMBMS.** Mandatory interface with n:1 mapping of eMBMS-U to NAS UE Data Layer for supporting the transfer of application data (GTPv1-U over UDP/IP).

Orchestrator input/output for function composition

- For further evaluation in WP5

Controller input/output for function management

- For further evaluation in WP5

6.21.2. NAS UE Specific and Control Layer

Description

NAS UE specific c-layer functional block refers to the user perspective functions and procedures related to the non-radio signalling between the UE and MME.

Details

We can identify different functional blocks:

- Connection establishment:
  - Registration-NW attachment and bearer management: this function allocates a default bearer, activate and deactivate bearers.
  - IPv6 neighbour discovery and IPv6 router discovery
  - NAS signalling to UE: this function conveys information between the UE and the core network. It is used to manage the establishment of communication sessions and for maintaining continuous communications with the user equipment as it moves.
- Security functions:
  - NAS security: this function handles all the ciphering, user identity confidentiality, and integrity protection and authentication operations.

  - Authentication: this function handles the user authentication, key agreement (exchange), public key generation.

- Mobility functions:
  - Tracking Area Management: this function manages the changes of the tracking area (TA).
  - Inter-node signalling for mobility: this function forward users packets while the S-GW is acting as the mobility anchor for the data layer during handover operations.
  - Context transfer function: this function is in charge of transferring the UE context from the source eNodeB to the target eNodeB during handover.
  - Inter-operator charging: this function accounts for packets processed by S-GW and data collection and usage per UE.
  - Paging: this function is used to convey messages to the UE switching from IDLE to CONNECTED mode. The UE decodes the content (Paging Cause) of the Paging message and initiates the appropriate procedure.

Interfaces

- **NAS UE Specific and Data Layer.** (1:n) exchange c-layer information in order to issue d-layer user functions
- NAS Event-Control Layer. (1:1)
- NAS Core Network. (1:1)
- **RRC Cell Specific.** (1:1) exchange c-layer information between cells and core network
- **RRC User Specific.** (1:1) exchange c-layer information of UEs

Orchestrator input/output for function composition

- For further evaluation in WP5.

Controller input/output for function management

- For further evaluation in WP5.

6.21.3. NAS Event-Control Layer

Description

NAS Event-C function block refers to the network-side c-layer functions and procedures including those of the interface between RAN and CN (S1-C in a LTE-like example [36.300]), which are provided to facilitate mobile connectivity for UE.

Details

NAS Even-C may be further divided into:

- UE radio connection management functions
  - 5G radio access bearer (RAB) service management: establishing, modifying and releasing 5G RAN resources for user data transport for a UE once a UE context is available at RAN with respects to network controlled QoS requirements and multi-connectivity modes. UE context in 5G may include contexts of application/device profiles. Note that 5G is expected to support ultra-high availability, ultra-high reliability and ultra-low latency services and massive MTC. Hence, 5G may adopt new or different bear service model, networking paradigm or QoS concept, as compared to LTE.

    - UE context management: establishing, modifying and releasing UE contexts for UE in CONNECTED state to a serving 5G network in order to support user individual signalling on the interface between the serving RAN and CN, also to enable and facilitate advanced UE context aware networking features in 5G.
- Mobility management functions
    - Reachability management primarily for UE in IDLE state if to be supported in 5G: main functions include tracking UE location on a granularity of one or multiple overlapping tracking areas, referred to as Tracking Area List in LTE, so that UE can be paged and therefore reached. UE needs to perform tracking area register and tracking area update on a regular basis, as preconfigured. Both UE and network sides adopt necessary timer mechanisms to keep and validate up-to-date tracking location and reachable state of UE. Tracking Area List is managed by the network by allocating and reallocating the Tracking Area Identity list to UE. All the tracking areas in a Tracking Area List to which a UE is registered are served by the same serving LTE-like MME. In a multi-RAT environment of 5G, UE may be reached via different RATs. Supports of ultra-high availability, ultra-high reliability and ultra-low latency services and massive MTC in 5G pose certain challenges for reachability of such users in terms of at least feasibility, manageability and efficiency.
    - UE location reporting from RAN: CN (LTE-like MME) may request serving RAN (LTE-like eNodeB) to report UE location information.
    - Mobility restriction management: functionality is provided by the UE, the radio access network and the core network, as configured and controlled by the core network for certain restrictions of UE mobility in RAN.
- CN serving node selection functions: for selecting, reselecting or relocating optimised serving network nodes (gateways or servers) in CN for UE, depending on how 5G architectures define such serving network nodes physically (if at all, considering a possibility of using common virtual network control entity). In LTE these functions include PDN GW selection, SGW selection, MME selection, SGSN selection and PCRF selection, as the c-layer is spread across many entities. In 5G, SDMN based c-layer may be considered as a single logical centralised control entity. However, the practical operational requirement of no-single-point network failure coupled with scalability requirements that are applied for PLMN in general imply some sorts of selection, related to flexible and scalable physical location of VNF for user as well as network function instances anyways. There can be many selection criteria and triggers for selecting, reselecting or relocating relevant CN serving node, including network topology, optimised mobility management, load balancing, optimised data-path handling, and so forth, as described in [23.401]. 5G NORMA considers adaptive optimisation from functional view, topological view, resource view and deployment view for a flexible multi-radio multi-service 5G network with SDMN driven control and orchestration. Hence, highly flexible and scalable CN serving node selection functions may be expected.

Interfaces

NAS messages and c-layer messages for NAS-AS interactions for UE are exchanged between NAS Event Control Layer and:

- **RRC Cell Specific**. 1:n for per cell common control for all relevant UEs if any
- **RRC User Specific**. 1:n for per UE control
- **NAS UE Specific and Data Layer**. 1:1 for relevant control procedures (within NAS)
- **NAS UE Specific and Control Layer**. 1:1 for relevant control procedures (within NAS)
- **NAS Core Network**. 1:1 for relevant control procedures (within NAS)
- **Self-Organizing Networks**. 1:n for SON based control procedures
- **eMBMS**. Mandatory interface with 1:n mapping of eMBMS-C to NAS Event-C for supporting the MBMS session management, including QoS parameter (GTPv2-C, SCTP over IP).

Orchestrator input/output for function composition

- For further evaluation in WP5 (UE contexts and service requests, serving cell contexts, etc.).

Controller input/output for function management

- For further evaluation in WP5.

6.21.4. NAS Core Network

Description

NAS CN refers to network management functions which are provided to support O&M functions of 5G CN.

Details

In the current LTE, network management functions, as described in [23.401], include:

- GTP-C signalling based load and overload control
- Load balancing between MMEs
- Load rebalancing between MMEs
- MME control of overload
- PDN GW control of overload

The highly anticipated network virtualisation and centralisation in 5G may change or redefine serving CN nodes and associated network management functions in 5G. However, requirements on network management functions including load balancing and overload control for 5G are increasing. Hence, load re-balancing and load migration in the context of various scaling issues of network function and services for serving a number of tenants and mobile UEs with flexible and optimal location or relocation, addition or removal of a network function instance are expected. Therefore, NAS CN related functions in 5G NORMA may need to address, e.g., optimised re-routing, re-association between NF instances, UE signalling overload, and so forth.

Interfaces

C-layer CN load control signalling including triggers for possible path change or path switch in multi-path routing with a relocation of some serving CN node/site for one or more connections of one or more UEs or related functional entities are expected between NAS CN and:

- Not directly to RAN AS but via other NAS function blocks and SON with different control granularities or resolutions (per hosting site, per node, per tenant, per UE, per connection, per service or any group or combination thereof)
- NAS UE Specific Control Layer. 1:n
- **NAS Event-Control Layer.** 1:1 for relevant control procedures (within NAS)
- **Self-Organizing Networks.** 1:n for SON based control procedures

Orchestrator input/output for function composition

- For further evaluation in WP5.

Controller input/output for function management

- For further evaluation in WP5

6.22. RRC Slice

Description

RRC Slice function block is introduced to enable RAN slicing with possible RAN slice specific customization and control in order to serve UE according to tenant specific QoS policies including security and mobility aspects. RRC Slice is a part of dedicated control layer of individual network slice.

Details

Details are ffs and will be included in D4.2.

Interfaces

- **RRC User.** 1:1 mapping for exchanging control information related to slice-specific dedicated data layer functionality.

Orchestrator input/output for function composition

- Details are ffs and will be included in D4.2.

Controller input/output for function management

- Details are ffs and will be included in D4.2.

6.23. Geolocation Database

Description

A function block that stores information linked to geolocation, and makes decisions based on that geolocation information. In order to encompass the requirements of regulators, for example, in maximizing the benefits of such capability (e.g., in realizing spectrum opportunities through LSA, TV white space, etc.) it is noted that this should be a separate functional block, despite in some scenarios encompassing some capabilities/characteristics that could fit into other functional blocks of the network orchestration.

Details

Virtualisation, such as is considered within NORMA, commonly requires precise geolocation information. For example, geolocation must be known in order to be able to minimise the propagation path in the context of very low-latency applications that are envisioned for 5G. One prominent example of this is the Tactile/Haptic Internet or Tactile/Haptic Communications. In this case, computation availability and associated RF components that could build virtualised radio access must be chosen based on their locations to minimise the propagation path. It is noted that spatial traffic load hence the spatial capacity requirement changes/moves based on real geographical location, not based on logical location within a network. Given this, geolocation databases (GDBs) in the definition and management of RAN virtualisation are of fundamental importance more generally, for a wide range of applications and use cases that NORMA considers. It is also noted that there are big regulatory moves towards (broadly) GDB-based spectrum usage and sharing (e.g., TV white space, Licensed-Shared Access (LSA), light licensing, etc.), with the GDBs hosted at the regulator, regulatory-authorised/certified entities, and within the entities that are sharing their spectrum, such as an operator sharing their spectrum with another operator through an LSA approach, for example. Such GDB capabilities in NORMA give a link to the vast additional spectrum and flexibility that is realised through such concepts.

Prominently, geolocation databases might also assist greatly in other areas, such as rendezvous. For example, they might provide mutual awareness between two devices in a heterogeneous networking scenario that they both support (but don't have enabled, so aren't mutually detectable) certain radio access capabilities, or could virtualise/create those capabilities.

Regarding the technical specifics of this functional block, it is noted that it might both operate in the c-layer and d-layer, depending on the application for which the geolocation database is used. Moreover, it is generally asynchronous, with a high latency tolerance in signalling information, although again there are uses conceivable in which such tolerance will be less. Its multiplicity depends on realisation, but likely there will be only one per operator/network.

Considering Interdependences with other RAN functions, it is noted that interdependencies are present with SON, given that such GDB functionality can help to optimise the network in a self-organised fashion. Interdependencies are also present with RRC Cell, RRC User and RRC mmW,

as such GDB functionality can assist with improved RRC based on geolocation. Interdependencies are present with MAC, as such GDB functionality can assist the MAC decisions on access to particular resources. Other interdependencies are also likely, such as with RAT selection.

Considering the information exchanged between the connected functions/entities for this function block, it is noted that there will be reporting of geolocation information from virtualised (or virtualisable) network elements and devices to the GDBs. There will also be reporting of uncertainty in geolocation information (e.g., +/- X number of metres with Y confidence limit) from virtualised (or virtualisable) network elements and devices to the GDBs. Moreover, reporting of technical capabilities from radio devices and virtualised (or virtualisable) network elements (e.g., remaining resources for virtualisation, RF capabilities) to the GDBs will be necessary. Based on decisions, reporting will be done of resource availabilities and/or instructions/policies from GDBs to virtualised (or virtualisable) network elements and radio devices. Finally, reporting/confirmation of chosen resources from network elements and radio devices to the GDBs will be done. It is noted, however, that all of these information exchanges are very much dependent on the particular application for which the GDB is used.

Regarding the functional placement of this block, it is noted that it is centralised, but also can also be distributed; some examples of GDBs are currently distributed in the cloud. It is most likely to be in centralised cloud, except for in cases where it is necessary to minimise latency in signalling with GDB, in which case will be in edge-cloud.

Considering the requirements of this functional entity, these are broadly reflected in the "description" above, but generally:

- Geolocation capabilities must exist on network elements and likely devices (depends on implementation/purpose).
- Database and processing functionality in a (likely) centralised element.
- Reconfigurable RF capabilities in order to maximise the potential of use of such geolocation information.
- Support for signalling between the GDB and virtualised network elements, and likely also devices.

Finally emphasizing the gain and cost for virtualisation/centralisation under this function block, it is noted that there is very significant gain for virtualisation through being able to properly structure virtualised networks/elements based on actual locations where capacity is needed, required propagation paths, e.g., for delay reasons, etc. Further, there is very significant Gain through being able to "link in" to GDB-based technologies that are making progress, such as LSA, TV white space and light licensing. This might be done for extra spectrum and flexibility. Finally, the cost of implementation is minimal, as GDBs will be simple storage and processing functions. Some cost can be assisted with ensuring or implementing geolocation capability where it is needed.

Interfaces

The following interfaces apply for this function block:

- GDBs to/from RRC UE, RRC Cell, RRC mmW, RRC Slice, QoS Control, SON, RAT/Link Selection, MAC UE, MAC CA, MAC Scheduling, PDCP U, SDMC.
- GDBs to/from other virtualised network elements in general for the purpose of managing their computational resource usages on/among hosting equipment.
- GDBs to/from NORMA centralised control, such as a NORMA network orchestrator (SDM-O).
- Optionally but strongly beneficially, to/from the regulator/regulatory control; potentially also the GDBs might be run or certified by the regulator. This can open up vast amounts of spectrum through novel regulatory paradigms.

Orchestrator input/output for function composition

Logical address of the function (physical address(es) of hosting entity(ies)), computational requirements of the function, link requirements for signalling. This will be refined based on further discussion in WP5.

Controller input/output for function management

Geolocation information, resource capabilities, resource requirements (communication resource, e.g., spectrum, as well as computational/storage resource requirements), resource choices/instructions/options, acknowledgment/confirmation messages on chosen resources, perhaps policies, others. This will be refined based on further discussion in WP5.

PART II: INNOVATIONS

7. 5G NORMA Innovations

7.1. User-Centric Connection Area

LTE networks were mainly designed for the transmission of broadband packet payloads, i.e., a large amount of data is sent over the air interface, followed by a long time without activity. Once a UE is attached and registered within the LTE network, its connection to CN is managed with the help of the RRC protocol, which is terminated in the eNodeB. With respect to the radio activity of the UE, there are two RRC states for the UE namely RRC-Connected and RRC-Idle. The UE remains in the RRC-Connected state during the active data flow in UL or DL. The location of the UE is known on cell level.

In LTE, the connection management is controlled with the help of a RRC timer, which detects the radio in-activity. LTE allows a wide range of timer values. The study in [HQG+12] suggests timer values around 10 s based on the measurements in a practical LTE network.

Today's smartphone applications are developed such that they maintain their connection with the server using keep-alive messages. Additionally, there are also notifications from the server side. This is generally known as background traffic [ZZG+13]. Our motivation for signalling minimisation for 5G stems from this type of traffic. The background traffic mostly carries the packet payload in the order of a few bytes. However, due to the connection oriented nature of the LTE network, the setting up of the connection involves more than 20 signalling messages to transmit few IP packets.

Consider a scenario where a user is engaged in a "WhatsApp" based messaging (with packet inter-arrival time of around 10 s [36.822]) and the timer is set to a value, which is below 10 secs, then there will be a frequent transition between the RRC-Connected and the RRC-Idle states. Following the expiry of the RRC timer, the network will switch the UE to RRC-Idle state. Since the user is still engaged in a WhatsApp session, it may generate an uplink packet right after it has been switched to the RRC-Idle state. For this purpose, a completely new random access procedure has to be carried out and a new radio bearer has to be setup, which is detailed Figure 7-1.

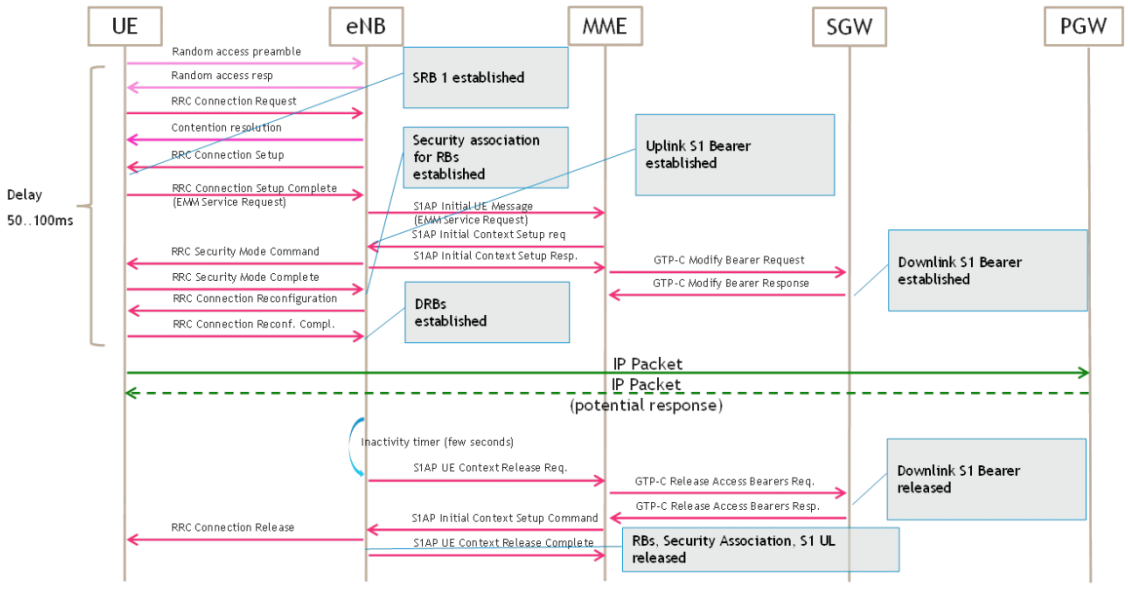


Figure 7-1: Connection Oriented Bearer Services and Small Data Transmission in LTE

On the other hand, the network may have a downlink packet for the user right after it has been switched to the RRC-Idle state. In this case, a paging procedure followed by a random access procedure is required. Typically, the paging is carried out within the complete tracking area where

many base stations are involved. This problem has been identified as "signalling storm", see [Sig_Storm]. Hence, we expect that in such cases, the given solution with RRC timer is not suitable enough to reduce the huge amount of signalling messages for small packet payloads on the air interface and towards the MME.

5G NORMA proposes a framework of a UCA for applications with small or sporadic data, which is depicted in detail within [ABA+16]. The main features and functionalities illustrating the advanced concept for 5G are listed in the following. The UCA framework relies on:

- The new 5G air interface supporting efficient asynchronous transmissions [SWC14] and smart protocols for small packets [SWS15],
- User centric connectivity architectures, edge cloud and virtualisation techniques,
- Ultra dense heterogeneous network deployments in 5G [GSA15].

The basic idea of the UCA is to dynamically allocate and update an anchor node within 5G RAN for each UE. The validity of the anchor spans over a user-centric coverage area consisting of one or more 5G-Nodes or radio cells. The anchor maintains the connection of the UE to the core network as long as it remains within the allocated coverage area by introducing a 5G RAN-controlled user-centric mobility. In the following, some functionalities of the UCA are described:

- **UCA definition:** Based on some user specific criteria (e.g. traffic profile and mobility), a centralised functionality defines the UCA for each UE. The UCA consists of a set of cells selected by a centralised control instance inside the 5G NORMA core cloud, see Figure 7-2. It can be defined based on the neighbour cell measurements but should include also anticipatory techniques, e.g. based on UE movement. The Core Network connections (bearers) are terminated at the anchor node, independently of the UE location within the UCA.
- **UE Context sharing within UCA:** The anchor node shares the user-context to the nodes within the UCA. Based on the context sharing, the UE is recognised in all of the cells within the UCA for transmission of UL packets and reception of DL packets.
- **Data transmission in UCA:** With the help of open loop synchronisation and efficient access protocols for 5G as described in [SWS15], the UE is able to perform both contention based and contention free UL transmissions in any cell (measured as best server by the UE) within UCA. The anchor node can configure the access methods that should be chosen by the UE with respect to the type of UL traffic and required service. For example, notifications for the best serving cell can be sent using a 1-step protocol [SWS15]. Small data can be sent using a 2-step protocol [SWS15]. The best serving node will forward the UL packets to the anchor node. The DL transmissions can also be handled with respect to the required service and the flow of traffic received by the anchor node from the CN. Assuming that the current location of the UE is known to the anchor node, e.g. with the help of best server updates, it will forward the packets to the current best serving node of the UE. The serving node can schedule the DL packet using a UCA specific UE ID. In case the uplink notification of a new best server was not received, the last serving node will not receive an uplink acknowledgment notification related to the downlink transmission. In this case, the last serving node will inform the anchor node, which will carry out a paging within the UCA. This paging is initiated by the anchor node, i.e., a RAN paging, and differs from the paging initiated by the MME.
- **UE mobility and UCA update:** A UE, which received a UCA definition, is allowed to move inside this UCA without an indication about the best serving cell, i.e., a UE based mobility is supported. Optionally, the UE might be configured to send a small data packet, i.e., an uplink notification about a new best serving cell. In this case, a backhaul link between the new best serving cell and the anchor node is established. This procedure avoids a lot of signalling messages inside the RAN and totally avoids update messages towards the MME.

Upon the detection of a radio cell not associated to UCA, the UE must send a measurement report to the RAN, which will carry out the UCA update procedure. This procedure is comparable to the handover in LTE, i.e., a measurement report indicating a new best radio cell and a list of other

radio cells in the vicinity is transmitted to the currently best radio cell of the UCA, which forwards the request to the anchor node. The new UCA will be defined by the central instance and will be communicated by an RRC message towards the UE and all radio cells of the new UCA. Within Figure 7-2 the procedure is explained, in which the UCA is updated as the UE moves to cell 8, which does not belong to the currently configured UCA.

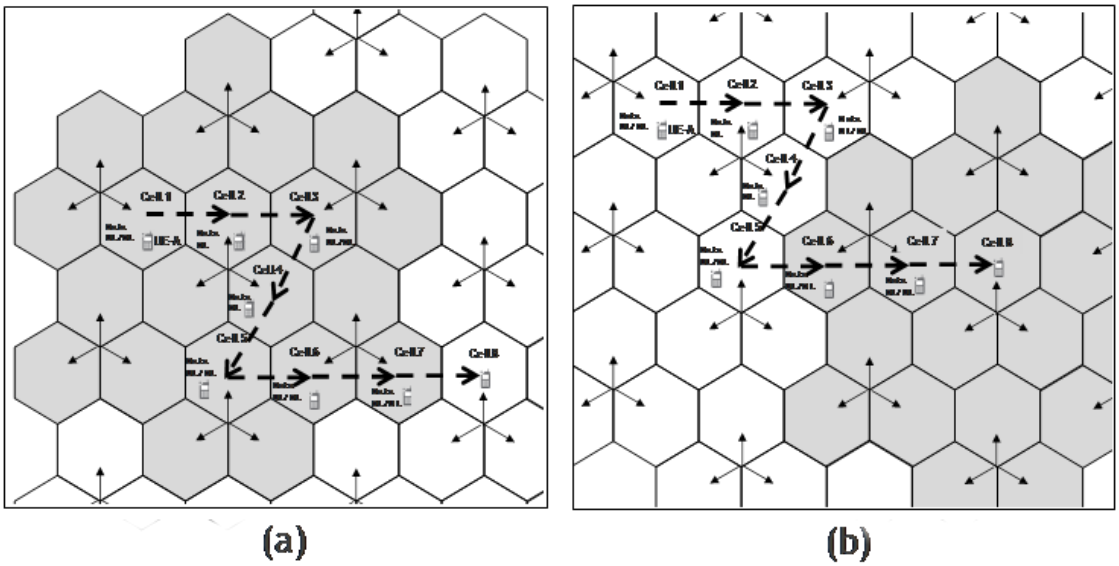


Figure 7-2: Assignment (a) and update (b) of MTA for a UE traversing the path shown by the dashed

New RRC sub-states: The configuration of UCA in 5G-RAN can be realised with the help of RRC protocol and the specification of new RRC sub-states. The UE remains in RRC-Connected state when it moves within the UCA. Therefore, no state transition signalling is required within UCA. With the help of RRC-Reconfiguration, the anchor node can enable or disable UCA for a UE as shown in Figure 7-3. When UCA is disabled, the RRC-Connected state can be seen as in 4G-LTE connected state. However, UCA is configured such that there is almost no or minimum signalling overhead on the radio interface as well as in the core network. Hence, UCA can be seen as a dormant state with low overhead but full connectivity for the UE both in UL and DL. Detailed information concerning the savings with respect to signalling messages can be found in [ABA+16].

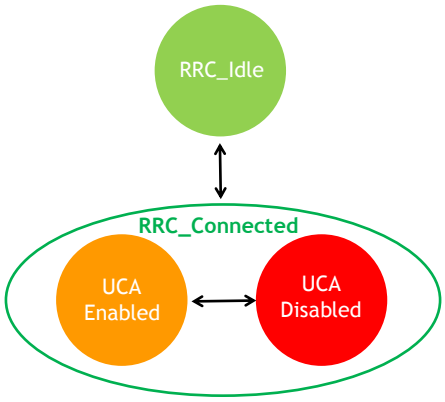


Figure 7-3: RRC configurations for UCA in 5G RRC-Connected state

7.2. RAN support for optimised on-demand adaptive network functions and services

7.2.1. Flexible on-demand configurations of RAN protocols

5G network architectures aim for flexible, adaptive decomposition, and allocation of mobile network functions and services on a per-service and per-scenario basis in order to optimise network utility and performance as well as end-user QoS and QoE. Therefore, 5G RAN architectures should enable and facilitate efficient support of different deployment scenarios and services with, for example, adaptive radio protocol stacks as shown in Figure 7-4.

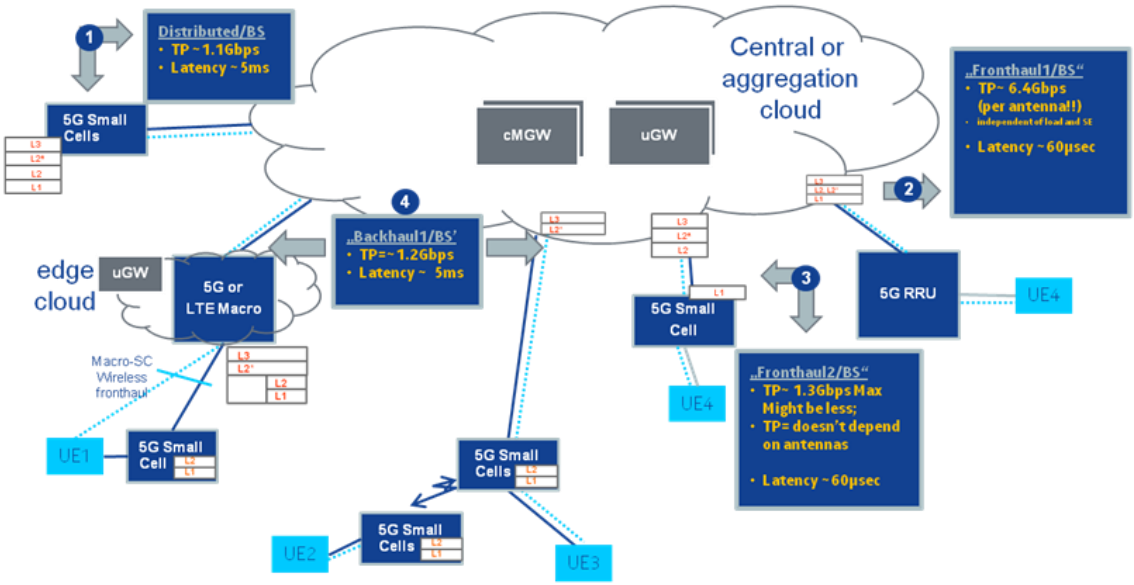


Figure 7-4: Example of flexible RAN protocol stack for different deployment scenarios

In such a flexible per-service per-scenario adaptive 5G network paradigm, the following observations can be made:

- Each 5G small-cell AP or eNodeB may have different capabilities to support functions and services across the radio protocol stack, e.g., low-cost APs may only support L1 and lower layer of L2 protocols but other APs targeted for stand-alone network deployment may support full radio protocol stack.
- The actual utilisation or effective use of RAN functions and services across the radio protocol stack provided by a 5G AP may vary, depending on network topology, available front/back-haul capacity, or requested user services. For example, a full radio protocol stack should better be configured and provided by an AP if local E2E services or services with low latency requirements are requested, while only some lower layer(s) of the radio protocol stack (PHY and MAC) may be configured and provided by the AP for serving delay-tolerant remote-access services in order to fully explore the advantages of cloud RAN.
- To serve different UEs with different requested 5G service flows (SF) and possible in-SF QoS differentiation (see Section 7.8 for more details on SF concept) in a flexible and optimal way may prefer different radio protocol stack configurations for different UEs even served by the same 5G AP or even same UE served by different 5G APs in case of multi-connectivity.
- The backhaul interface between 5G AP and the cloud needs to be open and capable of supporting flexible radio protocol stack configurations as well as multi-vendor inter-operability.

- Future network elements as well as functions may be implemented using, e.g., general purpose computing platforms, which allows for on-the-fly flexible customisation or enhancement of UE and RAN service capability with, e.g., on-the-fly on-demand software download and execution of an add-on piece.

In the current advanced cellular networks, standardised network elements have a rather clear, fixed allocation of network functions and services. Taking LTE for instance [36.300], the eNodeB is defined to provide all E-UTRAN functions and services plus implementing full radio protocol stack. There is an eNodeB configuration update procedure and an eNodeB may be configured and deployed in a certain SON based fashion but no flexible radio protocol stack related configuration has been supported in current cellular networks.

The iJOIN framework [IJOIN14-D52, iJOIN15-D53] investigated dense small-cell deployments with realistic backhaul limitations for targeted use-case scenarios including stadium, square, wide-area continuous coverage, shopping mall, and airport. iJOIN introduced the concept of "RAN as a Service" (RANaaS) to deploy functionalities partially or fully in a cloud platform, aiming for flexible and scalable centralised computing power as well as coordination gains. Although iJOIN seems to share a lot of conceptual elements with 5G architectures, especially from CN and RAN "cloudification" perspectives, it uses LTE as the targeted reference and does not address 5G challenging services and requirements (ultra-high reliability, low latency, IoT or massive M2M, high mobility V2X). In addition, iJOIN introduces RANaaS with flexible but clean enough RAN functionality split options, i.e, clean split but not dynamic distribution or redistribution of RAN functions and services across RAN protocol stack as per service and per serving scenario.

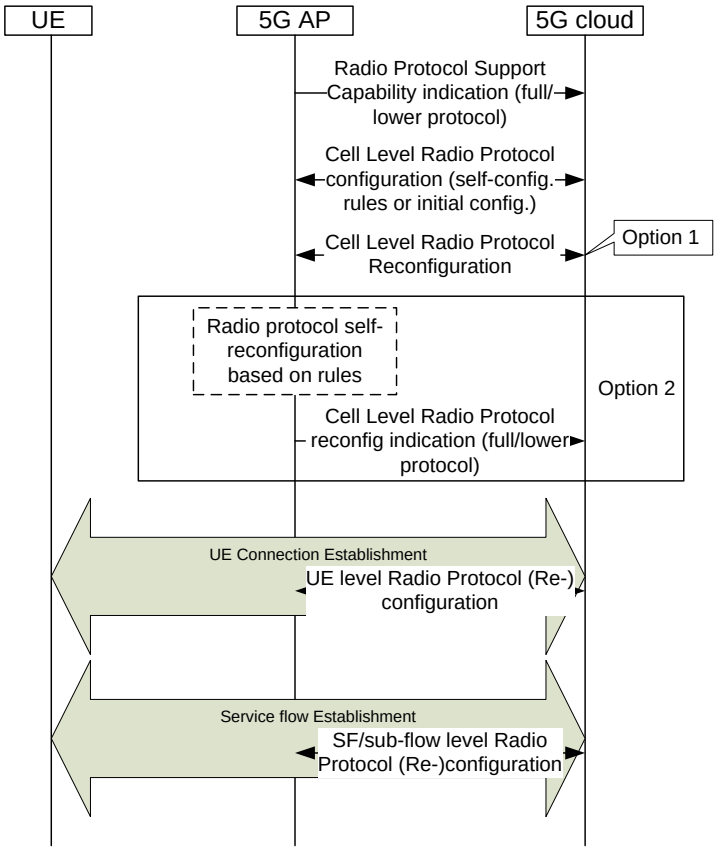


Figure 2.2-2: Signalling procedures for facilitating flexible radio protocol stack configuration

For enabling and facilitating flexible, on-demand adaptive configuration of RAN protocol stack, as described above, the following is proposed:

- Flexible RAN capability including generic hardware, software and front-haul capabilities to be exposed and managed dynamically on the fly,
- The RAN protocol-stack configuration herein refers to, e.g., an optimised placement (Cloud-RAN option) and configuration of each radio protocol layer specifically for each individual AP, UE, service flow or sub-flow of an UE (and not just radio bearer related parameter configuration as in current state of the arts).

Figure 2.2-2 illustrates some signalling procedures for implementing the proposed method. The detailed proposals includes a 5G AP, which during deployment, switch-on or (re-)activation, may indicate the capability of radio protocol stack support to the cloud where CN control entity or RAN aggregator (e.g. multi-controller) or SON and OAM server is located, e.g., information on whether higher layer protocol stack is supported or not, or whether in-bearer service-flow differentiation and application-aware scheduler is supported or not, open information on predefined general purpose computing platform or application interface which is not vendor sensitive information.

During operation, the 5G AP may be configured on-the-fly with the different radio protocol stack configuration mode even though the AP has full radio protocol stack capability. The configuration may be based on at least one of the followings:

- the cell load (e.g. in-bearer service flow differentiated scheduling in lower radio protocol stack may not be configured in low cell load case as high throughput and low latency can be expected for every service flows in this case);
- available front/back-haul capacity (e.g. in case of high front/back-haul capacity, AP may be configured to have lower layer protocol stack only so that advantage of cloud technology can be better explored by implementing higher layer protocol stack in the cloud);
- the user services served by the cell (e.g. full protocol stack may be configured if mainly local services with low latency requirement are requested by the UEs served in the cell.

Alternaively, the lower layer protocol stack may be configured in each involved APs if most of user services served in the cells request multi-connectivity so that the transmission on multi-connectivity can be more efficiently coordinated if higher layer protocol stack is located in the cloud). The radio protocol stack configuration of 5G AP may be from the cloud controller, CN control entity or RAN aggregator or O&M based on traffic monitoring or some report from 5G AP. Another option is that the 5G AP may be self-configured/-change the radio protocol stack based on pre-configured policies/rules. In this case, 5G AP needs to indicate the radio protocol stack configuration to the cloud every time when the configuration changes.

In some cases, the serving network may decide to customise or enhance service capability of a serving 5G AP as well as UE being served by the serving 5G AP with on-the-fly add or remove a piece of add-on software corresponding to certain network functions or services of the radio protocol stack. This is referred to as on-the-fly flexible configuration and control of software enabled radio access capability for both serving 5G AP and UE, which can be considered as a SON feature for 5G.

In case the full radio protocol stack is configured in 5G AP during operation, UE or service flow or sub-service flow specific radio protocol stack configuration may be performed when RRC connection is established or service flow or sub-flow is identified. The service flow or sub-flow specific radio protocol stack configuration may be based on service flow or sub-flow QoS. For instance, for the service flow with high throughput and high reliability requirement, multiple radio links from different APs may be established for transmission of the service flow. In this case, higher layer radio protocol may be configured in the cloud for more efficient coordination and control of the data transmission. To support the UE, service flow, or sub-flow level radio protocol configuration, radio protocol split request or indication signalling such as UE level radio protocol stack (re-)configuration procedure may be introduced between 5G AP and the cloud. The signalling may be embedded to the UE or service flow establishment message. As another option, in-

band signalling with service flow packet header masking or control-PDU may be introduced without control layer messages. The UE, service flow, or sub-flow specific radio protocol stack configuration in the network side may be either invisible or visible to the UE. For the latter case, UE AS and NAS procedure may be enhanced or adapted to the split radio protocol stack configuration (e.g. higher layer protocol is in the cloud and lower layer protocol is in the 5G AP).

SDMC Functions:

This innovation directly involves SON. SON on top of SDMC may provide all decision and determination of the proposed on-demand configuration of radio protocol stack for 5G AP.

7.2.2. On-demand RAN level decomposition of E2E connection

It is well known that for efficient utilisation of network and UE resources, the UE should consume just enough amount of resources only when it is really needed. This problem is addressed in many frameworks including RRM, QoS control, mobile context aware network optimisation, optimised support of smart phone users with always-on applications in mobile cellular networks, and efficient SON based ultra-dense small-cell networks. The operators wish to reduce power consumption, signalling and processing overhead while accommodating and serving as many users as possible for increasing revenue. The mobile users wish to have good QoS as well as prolonged mobile battery power or, that is, good QoE.

There are of course many techniques and solutions reported in literatures, addressing general as well as particular problems related to the above frameworks. However, practical solutions for advanced cellular networks so far often fall into L1 enhancing techniques, link adaptation on RAN level, on-off switching of mobile connection or some related physical resources, networking functions or services, SON based on-off switching of small cells.

In the related states of the art as mentioned above, it is often or traditionally assumed that the radio link is the capacity bottleneck of E2E mobile network connection. The reasons include, for example, scarcity of radio resources or unreliable radio channels. E2E aspects are often ignored when considering RAN problems or, on the other hand, RAN issues are over simplified when considering E2E problems.

Coming to 5G, it is expected that an active UE in a typical 5G small-cell deployment scenario is provided with an extremely fast and reliable radio connection which has a peak data rate of tens of Gigabit/s and hundreds of Megabit/s effective throughput or goodput on average. Thus, in most use cases, the backhaul connection between the serving RAN and CN provided to the UE may be the capacity bottleneck with an average data rate much smaller than that of the radio connection. It may be further assumed that 5G RAN may have certain service or application awareness or knowledge of higher-layer services or service flows of individual active UEs which are being served by 5G RAN. This service awareness may be at IP or higher layer(s), which may then be implemented as a part of enhanced QoS control features such as QoS-aware 5G PDCP or Uu application protocol of 5G RAN.

Let us consider the following use case scenario. Mr. Finn and their teenager son Junior are driving from Berlin to Munich. Junior wants to upload a large video file to Facebook using his mobile device. It may have potentially taken 1 minute for the UE to transmit the file to the serving 5G RAN but, said, 1 hour to upload it to Facebook due to the difference or imbalance in provided data rates of the fast 5G radio link and the much slower internet connection to Facebook server. The UE is in the meantime travelling across many serving cells.

The question now is how 5G supports for providing optimised on-demand service-aware adaptive network functions and services in 5G addresses such as the above use case with large data, long service session, high mobility and great capacity bottleneck problem from the CN part (end-user server). Note that E2E connection between a cellular UE and a remote user or server may cross many inter-connected administrative network domains. Those inter-connected network domains

other than the serving 5G radio access one may likely be able to provide moderate capabilities and capacities for the E2E connection, i.e., much less than that of the 5G radio access domain in serving the UE.

5G NORMA introduces a network functionality called UE agent based decomposition of E2E connections. This functionality employs an UE agent and decomposes or splits an E2E connection of the UE into two E2E connections: the first is between the UE and the UE agent located in RAN or inside a RAN proxy server; and the second is between the UE agent and the remote E2E server. This is in order to enable an optimised utilisation of 5G RAN capability and capacity for providing optimised QoS and QoE in serving the aforementioned use cases.

In particular, this functionality includes RAN level E2E decomposition of UP transport connection is applied depending on service of UE, meaning that service specific E2E transport connection between UE and, e.g., a remote server may be divided into 2 E2E transport connections: (i) the first one is between UE and the local UE agent; and (ii) the second one is between the local UE agent and the remote server. The local UE agent is considered as of RAN level which can be implemented or integrated in either a serving AP or a local server being close and connected to the serving RAN. The main purpose of adopting the local UE agent is to allow: (i) flexible adaptation toward CN in order to maintain best possible network connection to UE for the targeted service; and (ii) flexible adaptation toward 5G RAN in order to fully and efficiently utilise 5G RAN resources and capacity potentials as well as UE battery power. Thus, end-user content delivery and QoE for targeted services can be notably enhanced. The proposed RAN level E2E decomposition is determined and controlled by the serving network for specific targeted service of individual UE which may or may not be visible to the UE. The UE agent may or may not be aware of end-user application contents, which may or may not be visible to the UE as well. Figure 7-5 shows an example of the introduced RAN level E2E decomposition of UP transport connection.

In the use case scenario described above, the UE agent may be activated for providing and optimizing the upload service for the UE, allowing the UE to send the file to the UE agent in 1 minute and leave the UE agent to finish the upload in 1 hour. In the meantime, the UE may be put into an efficient power saving mode on RAN level waiting for a final service confirmation. In another example, considering a downloading case, the UE agent is activated to catch and store a large enough amount of data from an internet server without slowing down the internet connection (e.g., sending acknowledgement as soon as possible for advancing TCP traffic). Then the UE agent delivers the stored data to the UE over 5G RAN in a quick blast of an ultra-high data rate that fully utilises 5G RAN capability and capacity on offer. In this way, the UE does not have to stay in active state in 5G RAN for hour to download the file of a large content but only in certain occasions for a really short time periods (seconds) instead.

HO of CP and UP connections of an individual active UE may be handled flexibly, either together hand-in-hand as in current cellular networks such as 3G and LTE or separately in time for individual services or service flows depending on mobility and service characteristics of the UE for efficient utilisation of 5G RAN and UE resources. That is, actual HO of UP connection for an active UE may be treated on individual service or service flow, triggered depending on also the contexts of activated UE agent thereof. Thus, the current serving AP of the UE for at least the CP connection may be different from the AP which is serving or operating an activated UE agent for a targeted ongoing service of the UE. This is considered as one step further compared to prior arts in providing on-demand service and mobility.

In the uploading case, the UE agent is activated to get the file transmitted to it from the UE when being served by AP#k. The actual upload of the file to Facebook is handled by the UE agent and fully completed in 1 hour and at that time the UE has moved to the cell area of AP#n. The uploading session of the UE may be considered as ongoing all along the road from AP#k to AP#n, passing a number of other APs at some of which the UE may need to re-establish the CP connection (handed over) in order to update its UE contexts. The innovation herein allows the network to re-establish or hand over the UP connection of the UE only at AP#n when it is needed for the

UE to receive the final confirmation and not at other APs along the road when the CP connection may be handed over or re-established from time to time. In the downloading case, the UP connection for the UE needs to be re-established or handed over to the current serving AP only when the activated UE agent has downloaded a large enough amount of data for the requested content and determine to deliver that part to the UE. Thus, the need of HO or reestablishment of the UP connection for the UE may be triggered from the UE agent based on monitoring the progress of the UP connection toward CN. 1:1 mapping between the context of the activated UE agent and the up-to-date active/idle mobile contexts of the UE needs to be maintained at least in the network side.

Explicit enhancement of end-user QoE is enabled, as the activation of the UE agent for the targeted service of the UE can be made aware to the end-user and let the end-user to decide if it wants to delegate the upload or download task to the UE agent of the network. For instance, the serving network may issue a notification message to the UE that the internet connection is slow and it should be faster and more efficient for the UE to let the network assist in downloading or uploading large contents. The UE may request for activating the UE agent mode for at least one targeted service and then follow explicit control of the serving network.

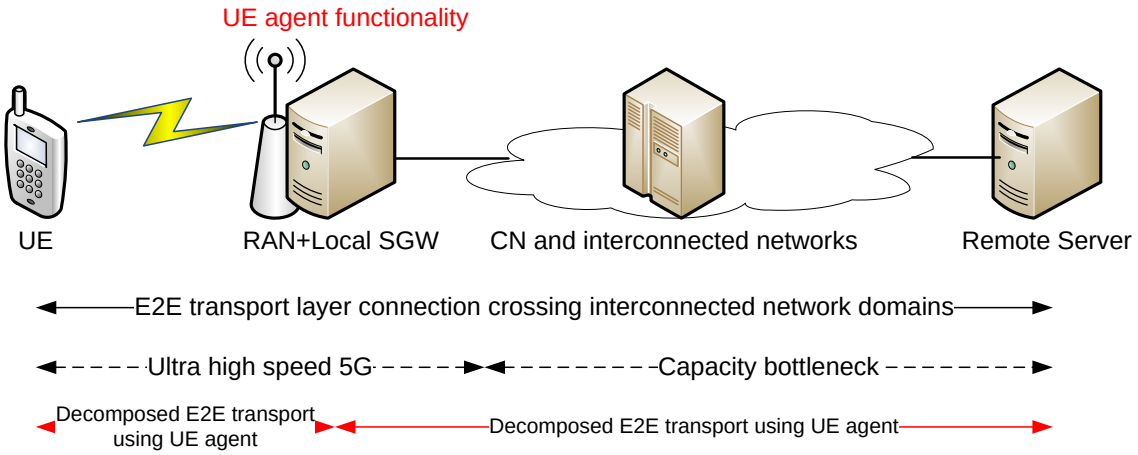


Figure 7-5: E2E data layer transport connection decomposition

Figure 7-6illustrates a signalling procedure for setting up the UE agent based decomposition of the E2E transport connection on-demand. In this example, full network control is assumed or, that is, the setup of the UE agent and use of the UE agent-based decomposition is initiated by the serving RAN side with minimum awareness from the UE side.

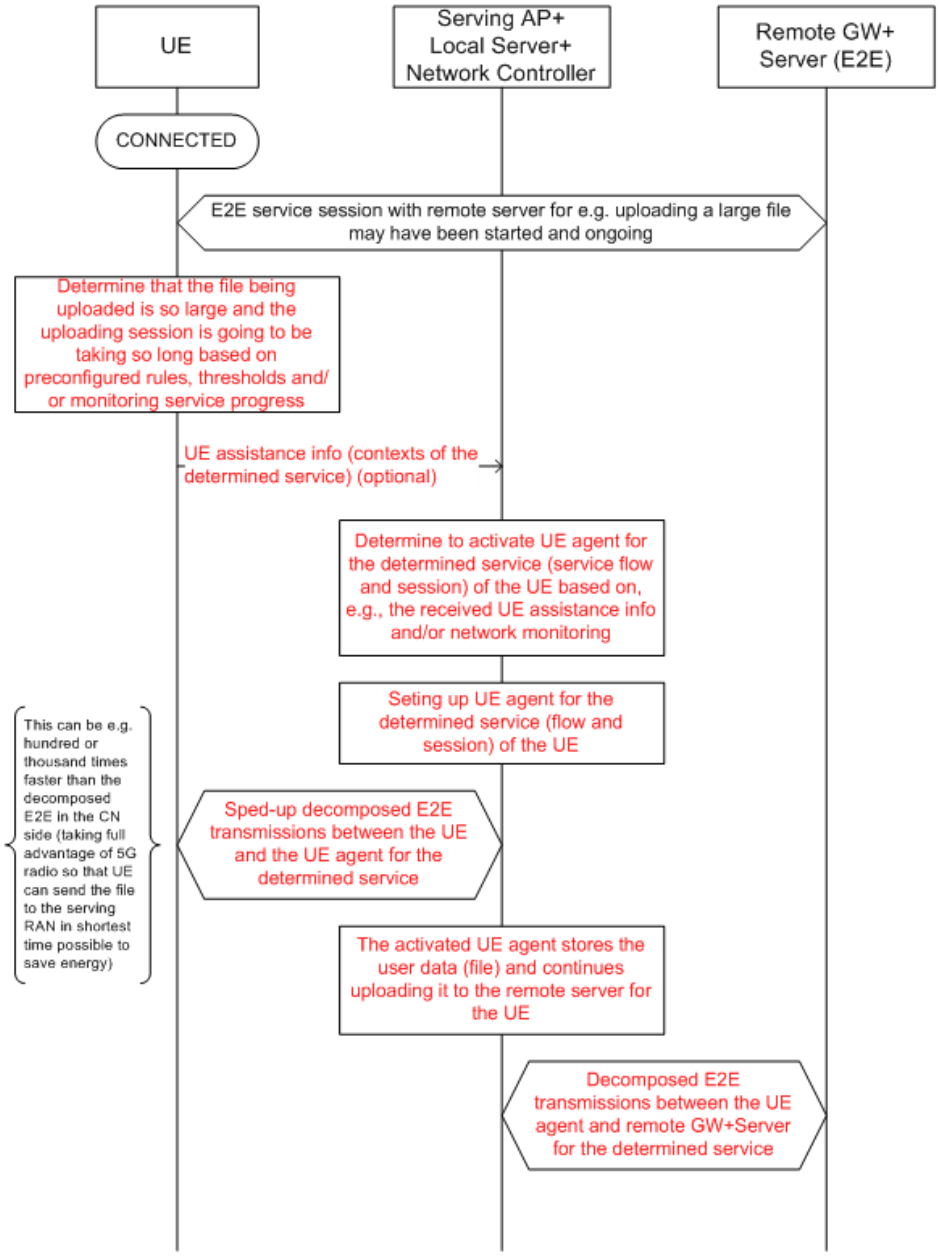


Figure 7-6: Illustration of UE agent based E2E decomposition setup and operation

Figure 7-7illustrates an on-demand HO with possible separation of HO for CP connection and HO of UP connection for the UE served with an active UE agent.

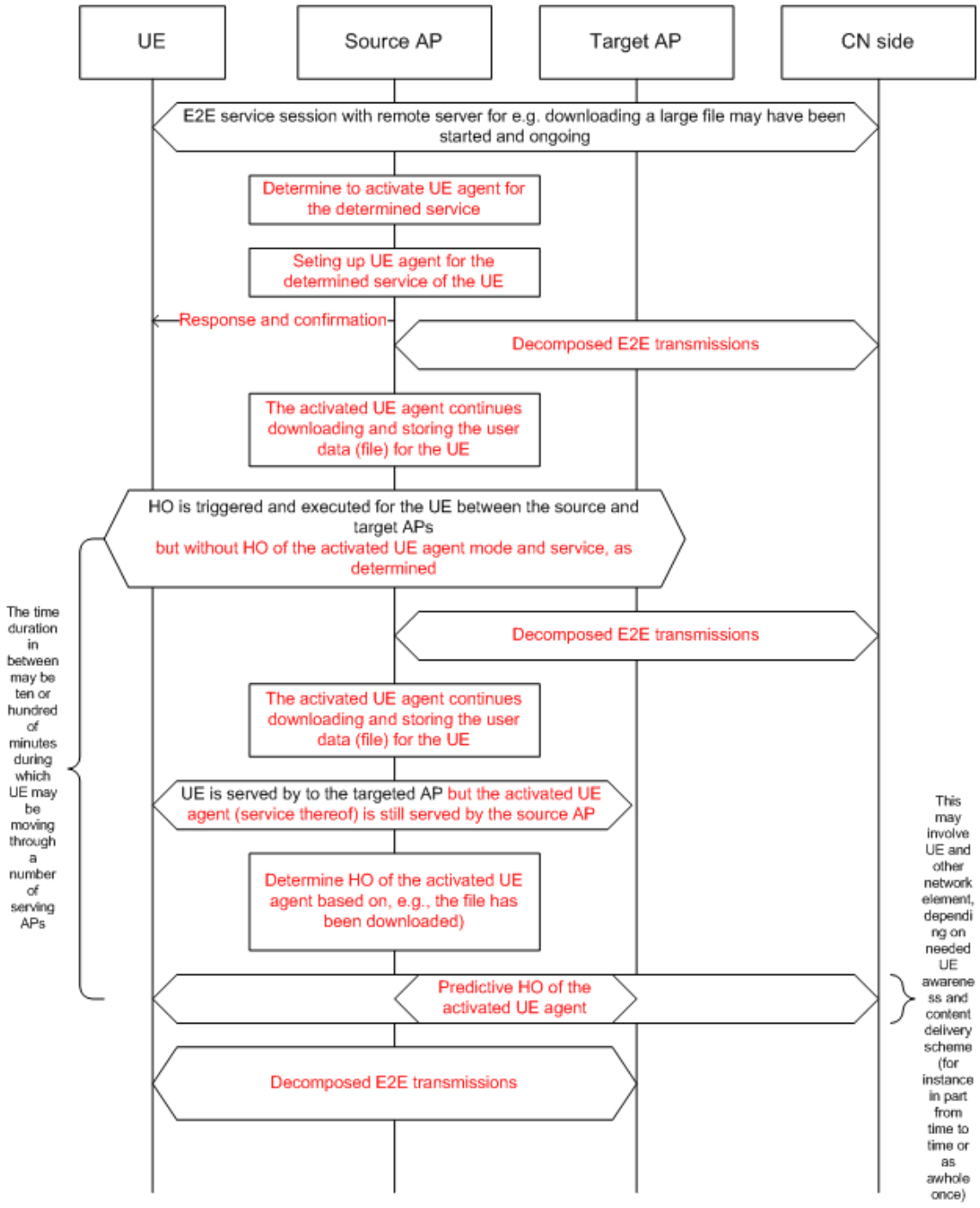


Figure 7-7: On-demand handover of UP connection with UE agent based E2E decomposition for signif-icantly reducing overhead

SDMC Functions:

SDMC NAS Control and QoS Control functions may interact with RRC User to facilitate this innovation directly.

7.2.3. RAN orchestrator for flexible RAN functions re-location

Problem:

In Section 4.1 and, in particular, Section 4.1.2 indicates that the common PDCP based MC solution has emerged as one of the most preferable options for MC functional architectures for 5G. In

this common PDCP based MC solution, PDCP functional entity in the serving RAN handles functions such as multi-connectivity anchoring or encryption. The serving network (RAN) hosts a PDCP entity per a RB or service-flow (SF) of a given UE, which may have one or multiple RBs or SFs. In cloud-based flexible RAN architecture, the network node or hosting site of PDCP for one or more UEs may be flexible (not necessarily in the serving primary macro eNodeB as in LTE. That is, physically, PDCP functions may be instantiated on generic transport network elements, or some virtualised/cloud-based computing platforms, or more conventional base-station platforms. For setting up RB service for an SF for a given UE, a suitable PDCP hosting site has to be selected for that UE. During the lifetime of an SF, PDCP relocation from a current hosting site to a new hosting site may be triggered for various reasons – e.g., processing overload, mobility impact, and/or E2E QoE expectations of certain sub-flows of a SF or bearer service.

Thus, the selection or reselection of a suitable PDCP hosting site for a SF or bearer service of UE requires awareness of various factors, which may not be available to RAN in current cellular networks.

5G NORMA introduces a new functional entity in the 5G RAN architecture, referred to as RAN orchestrator, which:

a) Receives registration notifications from one or more network elements when the network element hosts PDCP functions
b) Uses an SDM-C northbound API to obtain information about network topology (path, bandwidths, latencies, etc.)
c) Receives available capacity or load indications from the network elements regarding the PDCP functionality they are hosting
d) Receives a request from a network element regarding the need for initiating relocation of PDCP function or functional entity for one or more UEs or SFs or RBs thereof
e) Selects a target network element to host the PDCP function, considering transport network topology and state information, QoE/QoS requirements of corresponding SF or RB service of UE, and UE's multi-connectivity status
f) Notifies the requesting network element of the selected target network element to perform PDCP re-location

Logically, RAN orchestrator can be considered as a part of the RAN control layer or O&M. Thus, RAN orchestrator may be collocated with RRC User or SON function block in 5G RAN, interfacing with PDCP (RRC User and/or SON).

Figure 7-8 illustrates network functions under Steps a), b) and c) above.

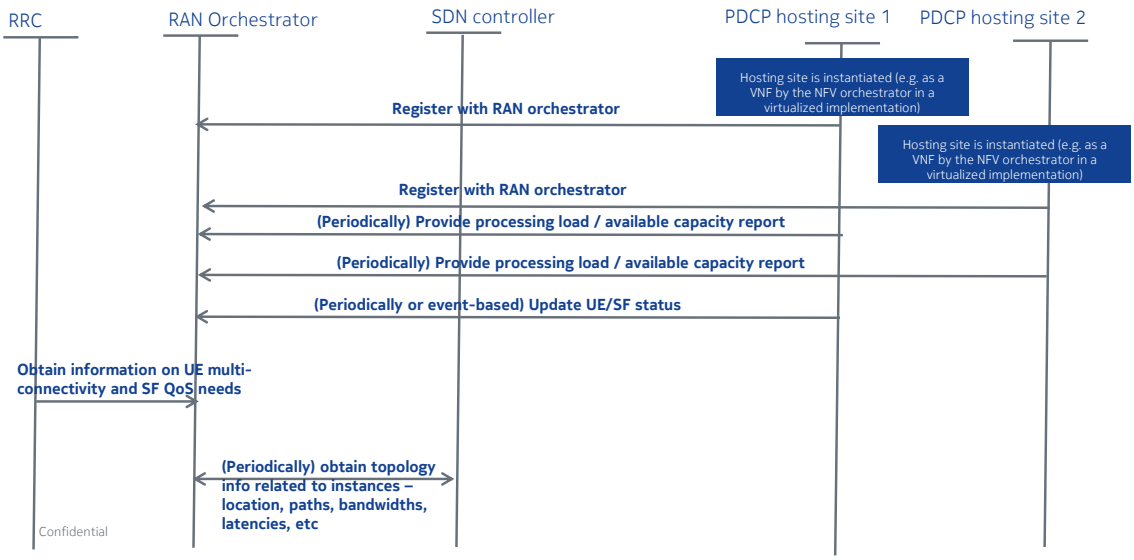


Figure 7-8: Obtaining awareness of PDCP hosting sites, load, transport network

Figure 7-9 illustrates some network functions under Steps d), e) and f) above.

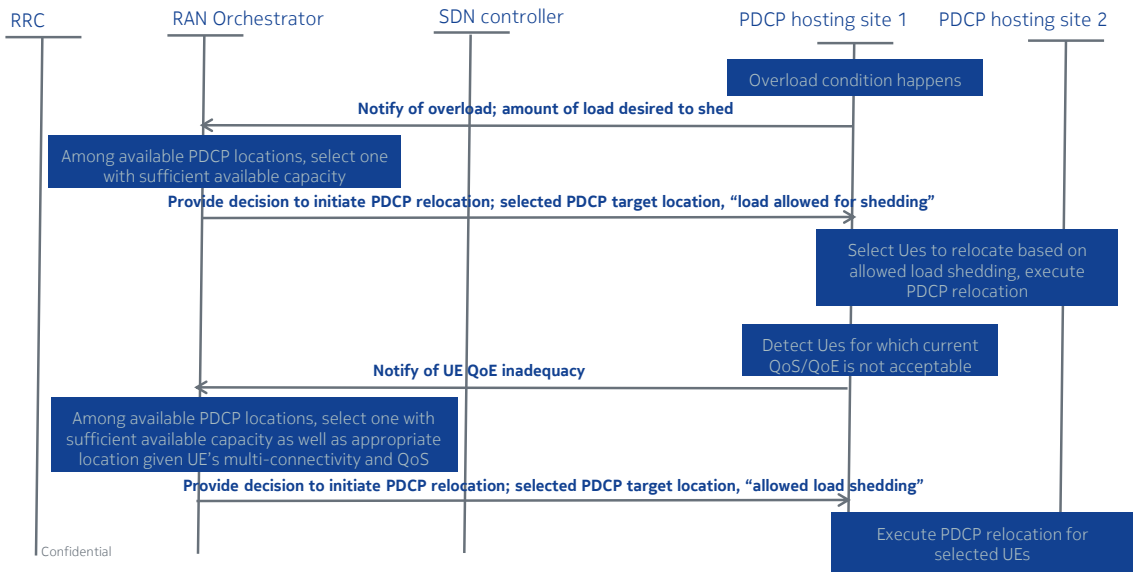


Figure 7-9: RAN orchestrator making decision to select target PDCP hosting location

Finally, Figure 7-10 illustrates an example of a decision-making process at RAN orchestrator for PDCP hosting site selection for SF setup or PDCP relocation for a UE.

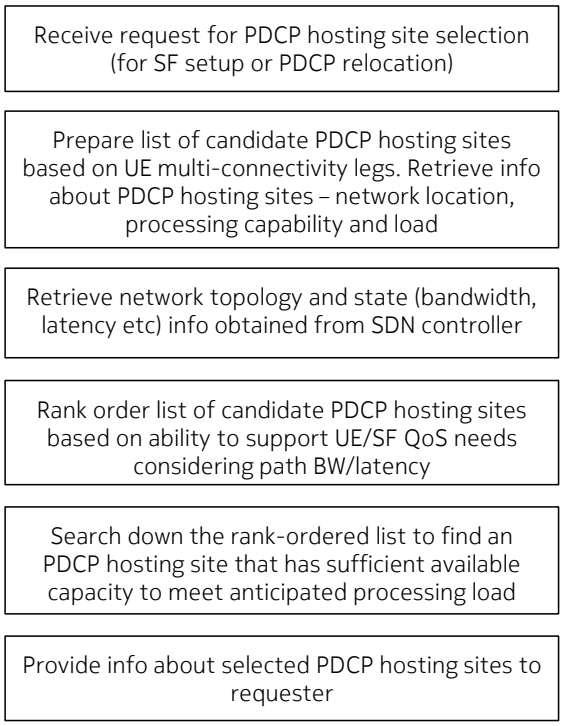


Figure 7-10: Example of a decision-making process at RAN orchestrator

SDMC Functions:

RAN orchestrator may be implemented as an extended part of SON. RAN orchestrator can be viewed as an application on top of SDM-C, which uses SDM-C northbound API to get information on RAN network topology (paths, bandwidths, latencies, etc.). Thus, RAN orchestrator can make decisions where awareness of the network topology or state (bandwidth, latency, paths) is needed. For example, in the common PDCP based multi-connectivity architectures described

in Section 4.1.2, when a given PDCP instance has to decide how to partition or split a given SF's traffic across multiple multi-connectivity legs, it can use transport network state information which RAN orchestrator has obtained from SDM-C. RAN orchestrator should also have a complete view on RAN level network topology – that means information on location of all RAN functions. Thus, in a flexible RAN architecture with flexible on-demand function decomposition, RAN functional entities or network elements such as small-cell APs, RRUs, and so forth should also register themselves with RAN orchestrator.

7.3. Mobile Edge Computing and Network Resource Allocation for Multi-Tenancy

Mobile network operators endeavour in a twofold mission that, on the one hand, is looking at enhancing traditional services (e.g., telephony, web and multimedia), and, on the other hand, aims at integrating in a single network infrastructure new vertical segments for public safety, healthcare, utilities management, connected vehicles and industrial automation [NGNM_WP]. Such technology leap is enabled by the virtualisation and softwarisation of the network infrastructure, which are pushing MNOs towards quicker network upgrades at lower costs, with the objective to build a flexible network infrastructure able to accommodate a plethora of diverse new services. In order to leverage the agile and elastic characteristics of cloud technology, the telco industry is working towards developing systems that will enable the cloudification of the existing network architecture and service provisioning.

Framed as an ETSI ISG, Mobile Edge Computing (MEC) focuses on evolving the mobile network's edge, in order to create a cloud-like environment close to the Radio Access Network that hosts enhanced services provided by the MNO or third parties. Such services span across caching for Content Delivery Network (CDN), RAN analytics, vehicular communications, IoT systems, benefiting applications a closer deployment to the User Equipment (UE) [ETSI_MEC02]. Such Mobile Edge (ME) applications run in form of virtualised objects on top of a generic cloud infrastructure located within the RAN, referred to as the Mobile Edge host. The Mobile Edge management system is responsible for managing both the infrastructure and the ME application instances that run on a single or different Mobile Edge hosts. Since the MEC management and orchestration system has operating characteristics similar to the NFV MANO, we argue that jointly orchestrating VNFs and MEC applications can provide several benefits from both the infrastructure cost and operation perspective, i.e. CAPEX and OPEX. The former is the ability of using edge cloud platforms commonly to support both applications and virtualised functions. The latter is the need of a common management and orchestration system as an enhanced version of the current NFV MANO, referred to as MANO+.

User scenarios lay the basis for the novel concept of network multi-tenancy, wherein the MEC concept plays as key-enabler [COMMAG_MT]. A Mobile edge platform is designed to offer enhanced services to Mobile edge applications, and the Radio Network Information Service (RNIS) is a key feature within MEC. Such service provides an API to ME applications to retrieve relevant information about the radio conditions on different metric basis (per user, group of users and so on).

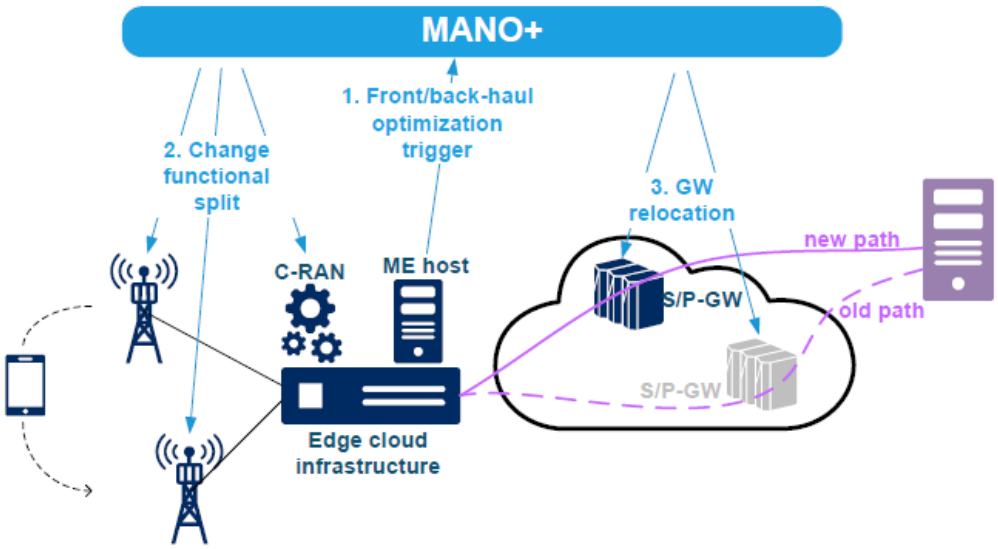


Figure 7-11: Operational flow for network slicing when MEC-NFV joint orchestration is in place

For instance, in a cloud RAN deployment where different functional splits may be applied, RNIS may be exploited by an ME application in order to compute the performance metrics of a given split. In our view, as depicted in Figure 7-11, the radio characteristics feed such an application that, in turn, triggers the MANO+ to improve the performance of attached users considering different functional splits. In particular, the MANO+ may examine alternatives in splitting the base station based on the performance resulting from application changes due to radio conditions or fronthaul dynamics. The MANO+ is required to apply a different functional split following two policies: i) increasing the capacity in the fronthaul by shifting RAN functions from the centralized Baseband Unit (BBU) towards the edge or Remote Radio Head (RRH), or ii) shifting base station functions from the RRH towards the BBU so as enabling cooperative multi-point (CoMP) schemes or better scheduling and interference coordination. Alternatively, a MEC fronthaul/backhaul optimiser function may trigger the MANO+ to provision changes in virtualised functions of the core network, including for example the re-location of a Serving/PDN-Gateway by shifting a virtual machine into a new location that guarantees a delay reduction or releases resources in the backhaul. This clearly sheds light on the reason why a novel MANO+ architecture is needed.

7.4. Multi-Service Technologies

The next generation of mobile networks is expected to enable a completely new class of services: Machine Type Communication (MTC), vehicle-to-X (V2X) communications, Internet of Things (IoT) traffic, device-to-device (D2D) communications, among others. Each new type of services will have their own latency, bandwidth and QoS requirements. An adaptable and flexible mobile network architecture is necessary to guarantee that these services are properly supported in 5G NORMA.

7.4.1. Current status of LTE

Even though LTE architecture was originally designed for voice and broadband services, it is quickly adapting to cope with the future requirements of 5G. The requirements of very high data rate, ultra-low end-end latency, extended battery life, multi-services and multi-connectivity have been addressed in the current 3GPP Releases. The investigation of different services and the respective technology used in LTE is presented in the below section.

Device-Device Communication and Public Safety Services: The 3GPP Release 12 addresses the two main services: Proximity Services (ProSe) and Group Communication. ProSe enable the

detection of other devices in the neighbourhood, allowing direct communication with PC5 interface between them [RS-12]. Direct communication is possible only in public safety scenario. Proximity Services work together with the EPC. The Direct Provisioning Function (DPF) is a part of ProSe that provides necessary parameters to the UE's for device discovery and communication. The ProSe function is also connected to Home Subscriber Server (HSS) to authenticate the device requesting direct discovery and configure it to ProSe policy [RS-12].

The direct discovery of proximal devices operates in two different modes:

1. Open Direct Discovery mode - UE's have access to discover all other UE's in the proximity
2. Restricted Direct Discovery mode – UE's do not have access to discover all other UE's in the proximity.

UEs are allowed to communicate with other UEs even in absence of network coverage. In case of special scenarios like public safety, no direct discovery procedure is necessary for communication between the ProSe devices. Three important scenarios are considered for the ProSe services:

1. Both the UE's have network coverage.
2. None of UE's have network coverage.
3. Either of the two UE's have network coverage.

The D2D communication in case of no network coverage is addressed by 3GPP Release 13 by allowing the UE in the network coverage to serve as a relay for the UE that is out of coverage.

Beside D2D, Group Communication and Mission Critical Push to Talk (MCPTT) (both public safety services) are introduced in Release 12. Group Communication Service Application Server (GCS AS) is responsible to select either unicast or broadcast mode to enable communication from UE to public safety devices. The MCPTT service is one-way communication that allows UE to talk and all others to listen. MCPTTP server is responsible for unicasting, multicasting or broadcasting the data to other devices in the group. The MCPTT service will be included in the Release 13 [NOKWP].

Massive Machine Type Communication: Due to tremendous increase in machine type communication in future, LTE focus on providing low cost and low power solutions. Support for low cost MTC devices is achieved by excluding MIMO and higher modulation schemes. Since, machine-to-machine communication can operate well with lower data rate, MTC devices will operate with 1.4 MHz bandwidth and half-duplex mode in 3GPP Release 13. In addition to that, LTE will provide Power Saving Mode (PSM) of devices, where a device is switched off for a given time period. The 3GPP Release 13 is planning to increase the DRX cycle duration from 2.56 sec to 2 minutes and restricting the data rate to 1Mbps [NOK-W] [ERI-W]. As the machine type communication mostly occurs indoor, shadowing effects are also addressed in Release 13, by allowing path loss of 155 dB [NOK-WPa].

Vehicular communication: Since LTE did not provide vehicular communication services in Release 12, 3GPP Release 13 plans to provide V2X communication service through LTE Radio infrastructure as LTE enables low latency, high throughput and D2D communication features.

Broadcast Services: LTE provides flexible broadcast and unicast transmission. Efficient broadcast services were provided with the launch of enhanced Multimedia Broadcast, Multicast Services (eMBMS) in Release 9. The efficiency of broadcast services is enhanced by use of multi-cell transmission with Single Frequency Network (SFN) that adds the received signal strength for the users at cell edge instead of causing interference [NOKWP].

7.4.2. Current research on 5G

How to better support multi-service in a mobile network architecture is a current area of research. In this section, some relevant proposed 5G architectures and current projects on 5G will be described.

A Software Defined Wireless Network is defined in [BOS+14], where mobile network SDN controllers provide a northbound interface to services. These services can influence how traffic is handled. The architecture is flexible because of the use of SDN controllers. Virtual operators can share the same physical resources, dynamically adapting their share based on changes in demand.

A new 5G control layer using SDN and NFV is proposed in [YVU+14]. RAN and CN control is unified and implemented as control applications running in an interworking set of hierarchical controllers. New functionalities could be implemented as multiple coordinating control applications at different levels. Controllers in UEs provide native device-to-device communication, and high-level network controllers provide service chaining.

Taking a look at the requirements 5G raises, the work in [TR+15] proposes an SDN-based plastic architecture for 5G networks. The network architecture consists of a unified control layer (across core and radio access networks) consisting of three logical controllers, and a clean-slate data layer. Network functions can be either centralised or run at the edge, based on the requirements of different services. Tunnelling and gateways were eliminated from the data layer, which now consists of forwarding paths set up by the cloud infrastructure.

In the METIS project [MET-D6.4], multi-service support was enable by two processes. First, five so called Horizontal Topics were identified: Device-to-Device communication (D2D), Massive Machine Communication (MMC), Moving Networks (MN), Ultra-Dense Networks (UDN) and Ultra Reliable Communication (URC). Network functionalities that different topics have in common were grouped into Building Blocks. The goal was to use these blocks to guide the overall system architecture, addressing the requirements of multiple HTs. METIS also decomposed the functional elements and network functions present in its technical work, creating a pool of different network functions that can be applied as necessary to network elements, depending on the use case or service.

7.4.3. Congestion control in Machine Type communication

Machine to machine (M2M) communication is a key enabler technology for the realisation of the Internet of Things. An M2M communication system consists of a large number of machine type communication (MTC) devices which can communicate with the MTC servers or other MTC devices to accomplish specific tasks. It is predicted that by 2020, there will be 12.5 billion MTC devices in the world. If all the activated devices try to access the base station or evolved node B (eNodeB) within a short interval, congestion would occur at the radio access network (RAN). Also, the MTC involves large number of low power devices. These devices generate more signalling overhead than the actual data. The two major location that are more prone to congestion are described below:

- At RAN level: If large number of MTC devices are connected to the same eNodeB and consequently use the same channels leading to high contention.
- At the Core Network:
    - MME: If the device is moving with high velocity, the frequent handover, connection establishment and maintenance will generate large signalling traffic. The signalling traffic further increases if large number of the low power devices are simultaneously generating frequent handovers.
    - S-GW and P-GW: All the traffic passes through these gateways in LTE therefore creating congestion in the network.

Based on the location of congestion, the congestion can be classified into two classes as at the data layer and at the control layer. The d-layer congestion is caused due to large number of devices sending data in the uplink. Even if the data per MTC device is small, the total sum of all data can lead to congestion at the S-GWs or P-GWs. The control layer traffic is signalling traffic generated due to tracking area update, mobility management, paging or event specific trigger.

In order to avoid the congestion, different approaches are proposed. One of the approaches is access class barring (ACB). The ACB factor allows MTC devices to transmit their connection

requests with different probabilities. In some research work a proposal has been made to separate the RACHs used by H2H/H2M and M2M communications to avoid H2H users of being blocked from accessing the network in the presence of bursty M2M traffic. Also, the ACB factor and the timing advance are jointly used to reduce the RAN congestion. An algorithm is proposed to adaptively update the ACB factor. The update is dependent on the number of connection request that were served and the collision rate. This algorithm uses the fact that by increasing the ACB factor p, the number of transmitted preambles and the number of collisions are increased. On this basis, if the number of the observed collisions (i.e., simultaneous preamble transmissions) is more than a threshold, then the algorithm will decrease the ACB factor p for a certain value. On the other hand, if the number of successful transmissions is more than a threshold, then the ACB factor p will be increased. Also, 3GPP Release 12 has reduced signalling due of TAUs from MTC devices by increasing the TAU period timer.

In our approach, we propose the use of dedicated small cells that will offload the traffic of selective MTC devices that generate very high signalling overhead. Use of small cell will offload most of the traffic from macro cells hence avoiding the congestion on the cellular radio access channel. Furthermore, to avoid c-layer congestion, we propose grouping of the MTC devices that has the similar feature using the classification algorithms. The MTC devices can be classified based on the velocity (direction and speed both), the frequency of the uplink in case of stationary devices, amount and type of signalling data, priority etc. The signalling can then be significantly reduced by sending a common signalling message to the group of devices exhibiting the same properties. The focus of group formation will be considering the MTC devices with high velocity that introduce very high signalling traffic required due to frequent handover management. Once the groups are formed, the MTC devices can then be configured with the access class barring probability and the corresponding silent period (DRX).

7.5. Multi-tenant dynamic resource allocation

7.5.1. State of the art

Driven by the capacity requirements forecasted for future mobile networks as well as the decreasing margins operators are able to obtain, infrastructure sharing is emerging as a key business model for mobile operators to reduce the deployment and operational costs involved in initial roll-out (capital expenditure (CAPEX) and operational expenditure (OPEX) of their networks).

The issue of dynamically sharing resources between operators has recently received substantial attention both from industry and standardisation as well as in the research community.

Network sharing solutions are already available, standardised, and partially used in some mobile carrier networks. These solutions can be divided into passive and active network sharing: passive sharing refers to the reuse of components such as physical sites, tower masts, cabling, cabinets, power supply, air-conditioning, and so on; active sharing refers to the reuse of backhaul, base stations, and antenna systems, and it's labelled as active radio access network (RAN) sharing. However, these sharing concepts are based on fixed contractual agreements with mobile virtual network operators (MVNOs) on a course granularity basis (monthly/yearly).

3GPP has recognised the importance of supporting network sharing since Release 6, and defined a set of architectural requirements and technical specifications that have been continuously extended since then. The latest activities have focused on the definition of new sharing scenarios and requirements [22.101], and the corresponding network management architecture and functionality extensions towards on-demand capacity brokering [32.130].

Based on the enhanced capabilities of dynamic sharing, new business models for infrastructure owners are expected to emerge, resulting in new revenue sources. Indeed, such an approach supports not only classical players (mobile-operators) but also new ones such as Over-The-Top (OTT) service providers that may buy a share of a wireless network to ensure a satisfactory service for their users. Think, for instance, of Amazon Kindle support for downloading content from

anywhere, or paid-TV sports subscriptions including a premium for watching live games, to name just two examples.

Recent research efforts have also addressed the design of multi-operator sharing architectures. Building on eNodeB virtualisation, [K+12b], [CSG+13] introduce the notion of Network Virtualisation Substrate (NVS), a virtualisation technique that provides an interface for operators to reserve wireless resources. This solution is enhanced in [KMZ+13], which follows a gateway-level perspective to facilitate near-term adoption. Some other works in the literature have addressed this issue from an economical perspective [CCC+15], [F+08]. None of the above works deals with the design of specific algorithms for resource sharing.

In [GLK14], [MVA14], and [MKZ+13], the authors address the design of optimal algorithms for resource sharing among operators. In [GLK14], the optimisation of the total network utility is addressed by using max-min fairness; in contrast that relies on proportional fairness, which could provides many desirable properties. The works of [MVA14], [MKZ+13] present a proportional fair formulation however, they do not provide a rationale to justify their choice. Furthermore, [MVA14] does not address the design of an algorithm, while [MKZ+13] uses a general non-linear solver that incurs a very high computational complexity.

In a more general context of resource allocation in (single-operator) mobile networks, there have been some works in the literature addressing problem formulations [BLR06], [Y+13], [K+12b], [LPY08]. However, substantial attention has been devoted to the architectural framework for multi-tenancy, but relatively little work has focused on criteria/algorithms and state-of-the art ones fail to meet the requirements for a practical solution or they have been proposed without proper justification.

In designing a practical solution for dynamic resource sharing, however, we facing multiple challenges. To start, we need a sharing criterion that not only allocates resources to operators fairly, but also shares the resources of each operators fairly among its users. Furthermore, the criterion should allow for different pricing (or sharing) levels according to the operators' needs. When allocating resources to an operator, we should obviously take into account its numbers of users, but also their location, as there may be some locations where demand is higher and (consequently) resources more valuable. Given the amount of information involved (including ¡ the channel quality of each user) and its dynamic nature, the algorithm should be distributed. Also, since the algorithm may be triggered frequently (whenever a user joins, leaves or changes its location), it should be computationally efficient. When adapting to network changes, the algorithm should control the number of handoffs triggered, as those may represent a high overhead.

Hence we can conclude that there has been substantial work towards addressing this problem, most has focused on architectural issues, leaving algorithmic aspects open to consideration. A functional algorithm has to provide the functionalities and respect the requirements described above, in order to enable a multi-tenant networks.

7.5.2. 5G NORMA Contribution

Substantial attention has been devoted to the architectural framework for multi-tenancy, but relatively little work has focused on the design of criteria and algorithms for this purpose. While some algorithms and criteria have been proposed in the literature, these either fail to meet the requirements for a practical solution or rely on criteria that have been proposed without proper justification on their optimality.

In order to extend the Network Sharing 3GPP standardisation to meet the 5G requirements, our contribution involves the design of:

- New sharing criterion
- New sharing mechanism

7.5.3. New Sharing Criterion

3GPP standardisation provides a static allocation that guarantees a minimum level of resources and limits the maximum amount of resources allocated to a tenant. These resources could be available for a specified period of time and a certain location. In the context of 5G networks, our goal is to design a new sharing criterion that allows for allocating the resources among tenants in more flexible way.

The allocation of resources involves the following decisions: (i) the association of users to base station (where each user is associated with a single base station), and (ii) the sharing of the resources of each base stations among its associated users (the fraction of the base station's resources allocated to user).

We would like to allocate resources across operators dynamically, tracking changes in the numbers and locations of an operator's mobile users and changes in the associated transmission rates. When doing this we need to make sure that (i) network resources are fairly shared among the various operators, and (ii) at the same time, the resources allocated to a given operator are fairly shared among the users of that operator. A natural way to achieve this is to maximise the overall network utility resulting from aggregating operator utilities, defining the operator utility as the sum utility of the operator's users. In this way, maximizing the network's overall utility ($W(x,f)$) corresponds to maximizing the sum of operator's users utilities:

$$W(x,f) = \sum_{o \in O} \sum_{u \in u_0} \omega_u \log\left(r_u(x,f)\right) \qquad (7.5\text{-}1)$$

where:

- ω_u: the weight of user u,
- r_u: the relative throughput.

Furthermore, we need to consider the network share that is allocated to each operator (which captures the relative weight of the operator on the network). The shares might, for instance, be based on the level of (financial) contribution to the shared network: if an operator contributes twice as much as another, should roughly get twice the resources. Additionally, we also need to account for the number of an operator's users: if an operator has twice the share of another one, but also the twice as many users, its users should not be better off. To that end, we insert in the network utility the variable (w_u) that consider the network share of an operator, divided equally amongst its current users.

Furthermore, the new criterion has to allow the operators to allocate its network resources implementing specific policies, for instance, an operator may distribute the resources it has been assigned among its users prioritizing some users based on their traffic type, or may apply admission control to limit the number of users it has in the network, thereby providing a better quality to admitted users. This is an important feature from a business model perspective, as it allows an operator to adapt to the specific needs of its users and thus differentiate itself from competing operators.

The user utility function considered is particularly suitable for elastic traffic, the proposed approach can also be used when resources are shared with inelastic traffic. We envisage two alternatives: in the first one, we first run an algorithm to determine the resources needed by each operator to satisfy inelastic traffic needs, and then the sharing algorithm is used to share the remaining resources among elastic traffic; in the second alternative, the sharing algorithm is used to determine the resources provided to each operator, which then uses these resources to serve both inelastic and elastic traffic. Besides, we want to provide to network operators, a way to maximise their revenue deciding whether accept or reject an incoming inelastic/elastic new request.

Note that with this criterion we don't need to reserve some resources for tenants also when they don't need it and they don't use it, and the free resources could be available for the other tenants.

Furthermore, the tenants don't need to ask, like in 3GPP standardisation, for more resources because our criterion just provide of each tenant how many resources requested. Obviously if the resources requested are not available we need to respect a given outage probability.

7.5.4. New sharing mechanism

In order to implement the criterion designed above, we need to devise a novel mechanism for sharing resources. To this end, we focus on a dynamic sharing concept, signalling-based and with no human intervention, which enables a more efficient sharing of the network resources according to the agreed SLA and taking into account also the commercial requirements.

Given the amount of information involved (including the channel quality of each user) and its dynamic nature, the algorithm should be distributed. Also, since the algorithm may be triggered frequently (whenever a user joins, leaves or changes its location), it should be computationally efficient. When adapting to network changes, the algorithm should control the number of handoffs triggered, as those may represent high overhead.

The key idea of the algorithm is that, upon a new user joining the network, we re-associate a number of users and reallocate the resources of each base station to its users. To design such an algorithm, we first need to answer the following questions:

1. Do we really need to re-associate users?
2. Where should be (re)associated to?
3. In which order should users be re-associated?
4. How many re-associations do we need?

To answer the above questions, in the following we present some of the algorithm features and described how they should work at high level. The algorithm needs to consider the following cases: (i) a user joins the network, (ii) leaves, or (iii) changes her location.

In the optimal allocation, users are somehow balanced among base stations. When a new user joins the network, she greedily joins the base station providing the largest throughput. Now the balance is broken, so we may need to consider triggering user re-associations. The base station with which the user associates may have too many users. Hence, in the first step the algorithm has to re-associate one of the users of this base station. In the next step, the base station that received the re-associated user may have too many users; however, depending on the weights of the joining and re-associated users, the original base station may still have too many users as well. In each of these steps, we could select to re-associate the user that achieves the highest gain in terms of throughput. We repeat this, considering users from two base stations, in the subsequent steps. To limit their number and associated handoffs overheads we limit these to at most m. For the first m-1 re-associations, users choose the base station that provides the largest throughput, but in the final step (m^{th}) to avoid that the re-association of a user harms the overall performance, we select the base station association that maximises the overall network utility rather than the throughput of the re-associated user.

When a user leaves the network, the algorithm is quite similar: the base station with which the user was associated now may have less users than it could handle. Hence some other users may obtain a gain to associate with this base station. Also in this case we re-associate the user that obtains the highest gain. Now the other base station may have fewer users; we repeat this again for m-1 steps. In the final step we could decide again to re-associate the user that maximise the overall network utility.

When a user moves, her transmission rate values to the neighbouring base stations may change; if, as a result of these changes, at some point the user would receive a larger throughput in a new base station, we could re-associate her to this base station. Then, the old base station executes the same algorithm as when a user leaves the network while the new base station executes the algorithm corresponding to a joining user.

We have two possible ways to implement this algorithm:

1. We devise a multiplexer function with a scheduler per each network slices;
2. We execute a single scheduler for the network that allocates resources among different slices.

In the second case, we do not need a new function because we could implement directly the algorithm in the scheduler. In the first case the schedulers have to cooperate in order to obtain an optimal allocation of the resources overall network.

The new mechanism has to work also in Multi-RAT scenarios. In this case, we need to consider more possible allocation solutions because we share not only the resources of each BSs but we can decide to share also the different technologies of each BSs among all users. The problem now is obviously more complex.

As a conclusion, state-of-the-art efforts to date concerning multi-tenancy are more focused on architectural and requirements aspects, and less effort has been devoted to algorithmic aspects. To fill this gap, we aim at designing a completely new scheduling/multiplexing functions with new algorithm to allow the new 5G functions both for multi-tenancy in normal RAN and multi-RAT scenarios.

7.5.5. Proposed algorithms

In order to design new algorithms for multi-tenant approaches we have focused on 2 different scenarios:

1. Algorithm for handling resources requests at the infrastructure provider;
2. Algorithm for dynamic resource sharing among operators.

7.5.6. Admission control

One of the key novel concepts of the 5G architecture is Network Slicing: the infrastructure can be divided in different *slices* each of which can provide different services. This opens the mobile network ecosystem to new players:

- Infrastructure Provider (InP), which is the owner of the infrastructure;
- Tenants, which acquire a slice from the InP to deliver a specific service.

In this new ecosystem, tenants issue to the InP spectrum and computational resources requests in order to set up their slices. Since spectrum is a scarce resource for which overprovisioning is not possible and its availability heavily depend on SLAs and users' mobility, the InP cannot apply an "always accept" strategy for all the incoming requests. Thus, the new 5G ecosystem calls for novel algorithms and solutions for the allocation of network resources among different tenants.

Our idea is to design a network capacity brokering algorithm executed by the InP in order to decide whether to accept/reject a request from a tenant with the goal of maximizing the InP revenue, satisfying the service guarantees required.

In our model the InP receives requests from tenants characterised by:

- Amount of resources to be reserved
- Starting and end times for the reservation
- Type of traffic (elastic or inelastic) that imply:
    - Required quality/SLA
    - Its price ρ (amount of money per time)

The InP needs to decide which requests to accept knowing that by accepting a request with a small bid, it may lose future opportunities to involve higher bids (not enough resources available), but rejecting a request the InP loses the corresponding bid. Our algorithm leverage on Semi-Markov Decision Process (SMDP) theory, that models the resource allocation to network slices as a markov chain in which the next state depends only by the actual state, the decision taken and the transition probability function.

The algorithm requires the full knowledge of the system parameters like:

- the interarrival requests time λ;
- the request duration μ;
- the transition probability function;

and that the system is memory less. By applying decision theory, it is possible to find the decision policy that maximise the InP revenue. While SMDP provides the optimal policy, it requires very high computational cost as the space state is large so for practical purposes we need an adaptive algorithm, and we will use SMDP as a benchmark to evaluate the performance of the proposed algorithm.

The adaptive algorithm is based on Q-learning, a reinforcement learning tool that learns about the system behaviour by taking non-optimal decisions during the learning phase. After an initial learning phase, by evaluating the best possible action starting from a certain state, the algorithm is able to find the decision policy that maximises the InP revenue. The best advantages of this tool are:

- It does not need any knowledge of the system parameter (λ, μ and transition probability function)
- It works even if the system is not memory less
- It's an online algorithm that can react to system perturbations (it just need a short learning phase).

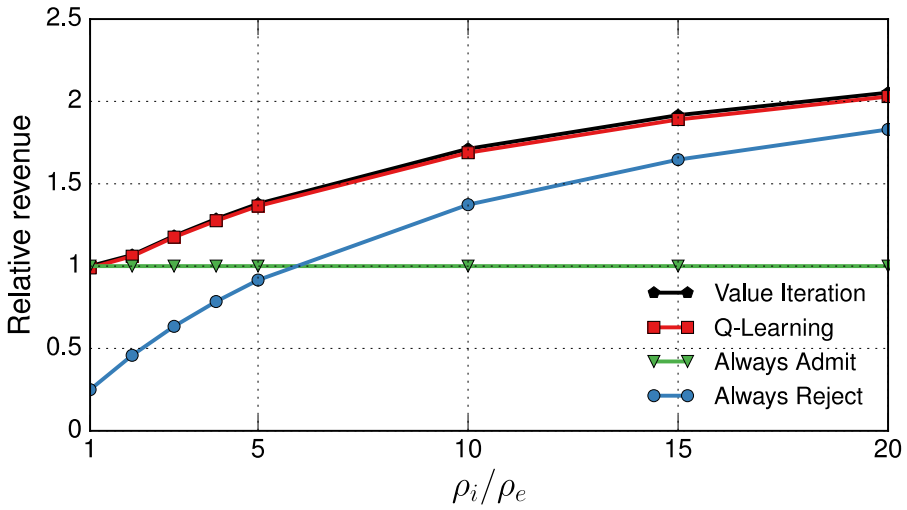


Figure 7-12: Relative revenue vs. ρ_i/ρ_e

In order to evaluate the adaptive algorithm, we compare the revenue obtained applying Q-learning policy changing ρ_i/ρ_e with:

- The one obtained applying SMDP approach
- The revenue obtained accepting all requests incoming (we reject them only if there are no resources available)
- The revenue obtained rejecting all elastic requests

We simulated two classes of incoming request: inelastic that demand a certain fixed throughput which needs to be always satisfied with a fixed outage probability; elastic that require an average throughput guarantees. Each class presents different parameters (λ, μ).

As we can conclude from Figure 7-12, SMDP always converges to the optimal policy, the one that provide the maximum revenue for the InP for each experiment. Furthermore, we can see that

with our adaptive algorithm we can obtain close to optimal performance even without the knowledge of all system parameters.

7.5.7. Dynamic resource sharing

Another important challenges in multi-tenancy are the definition of a sharing criterion and the design of an algorithm that follows it in order to enable statistical multiplexing of spatio-temporal traffic loads.

The idea is to design a criterion that maximise the network utility while:

- Allocates resources fairly among operators;
- Allocates the resources of each operators fairly among its users;
- Takes into account the number and the location of the active users of each operator.

The information involved include channel capacity, the number of users in the network, their spatial distributions and thus their mobility.

To be sure that our criterion achieves the above requirements, we define the network utility as the sum of operators' utilities weighted by the operator's share as:

$$W(\mathrm{x},\mathrm{f}) = \sum_{o \in O} s_o U_o(\mathrm{x},\mathrm{f}) \tag{7.5-2}$$

where:

- s_o :the network share assigned to each operator.

The operator utility is given by:

$$U_0(\mathrm{x},\mathrm{f}) = \frac{1}{|u_0|} \sum_{u \in u_o} \log\big(r_u(\mathrm{x},\mathrm{f})\big) \tag{7.5-3}$$

where:

- r_u: the throughput of user *u* of the operator *o*,
- U_o the set of users belonging to each operator.

With the above we can formulate the Multi-Operator Resource Allocation (MORA) optimisation problem as follows:

$$\max_{\mathrm{x,f}} W(\mathrm{x},\mathrm{f}) \coloneqq \sum_{o \in O} \sum_{u \in u_o} w_u \log\big(r_u(\mathrm{x},\mathrm{f})\big)$$

$$s.t: \begin{cases} r_u(\mathrm{x},\mathrm{f}) = \sum_{b \in B} f_{ub} x_{ub} c_{ub}, & \forall u \\ \sum_{b \in B} x_{ub} = 1, x_{ub} \in \{0,1\}, & \forall b,u \\ \sum_{u \in u_0} f_{ub} x_{ub}, \quad f_{ub} \geq 0, & \forall b,u \end{cases} \tag{7.5-4}$$

where the second equality and the last inequality correspond respectively to *user association* and *Base station resource allocation* constraints, and $w_u$is the user weights.

The proposed criterion allocates resources across operators dynamically, tracking changes in the numbers and locations of operators' mobile users and the associated transmission rates. Furthermore, the MORA criterion, satisfies some desirable properties both in the way base stations' resources are allocated to associated users, and the way users are associated with base stations:

1. Given fixed user associations, MORA allocates base station resources to the associated users proportionally to their weights;

2. The resulting resource allocation is Pareto-optimal, which means that if under some other user association choice, a user sees a higher throughput than that under MORA then there must be another user which sees a lower throughput allocation;
3. MORA is not harming any operator for the global benefit.

Compared with Static Slicing (SS) approach, where each operator contracts for a fixed slice of the network resources at each base station for its exclusive use, MORA provides a higher overall network utility and a higher operator utility for a given user association. For different user associations, there may be cases in which an operator sees a higher utility under SS than MORA, but the additional utility cannot be more than log(e). Another important result is represented by the capacity saving resulting from operators sharing infrastructure: sharing the infrastructure with MORA dynamic sharing provide a capacity saving that will be highest when infrastructure is shared by a large number of operators each with a small number of users per base station. With current trends toward small cells, the number of users per base station is expected to be small, suggesting that infrastructure sharing may be particularly beneficial.

The optimisation problem underlying MORA is a *non-linear integer programming problem*, which can be shown to be NP-hard, so an algorithm that provide the exact solution is not feasible.

Thus, we developed an approximation algorithm which performance are close to the optimal one, is semi-online (trigger a reassociation of a limited number of users upon a user joining, leaving or performing a handover) and distributed (due to the amount of information involved included the channel quality of each user).

In designing the algorithm, we need to decide:

- Where the users should be (re)associated
- In which order they should be reassociated
- How many reassociation are needed.

The proposed algorithm named Greedy Local Largest Gain (GLLG) is a modification of DG a simple distributed greedy algorithm that requires too many handovers and incurs too high overhead. In particular, with GLLG

- the re-association is done based on a largest gain policy, i.e., re-association are needed because using an online algorithm (upon a user joining the network, it only decides how to associate the new user, without triggering any re-associations of existing users) the performance can be arbitrarily bad
- the number of handovers is limited by a parameter m, i.e., in order to meet the best trade-off between the performance of the algorithm and reassociation overhead
- the eligible users to be reassociated is restricted locally to the ones within two base stations (the ones involved in the previous reassociation).

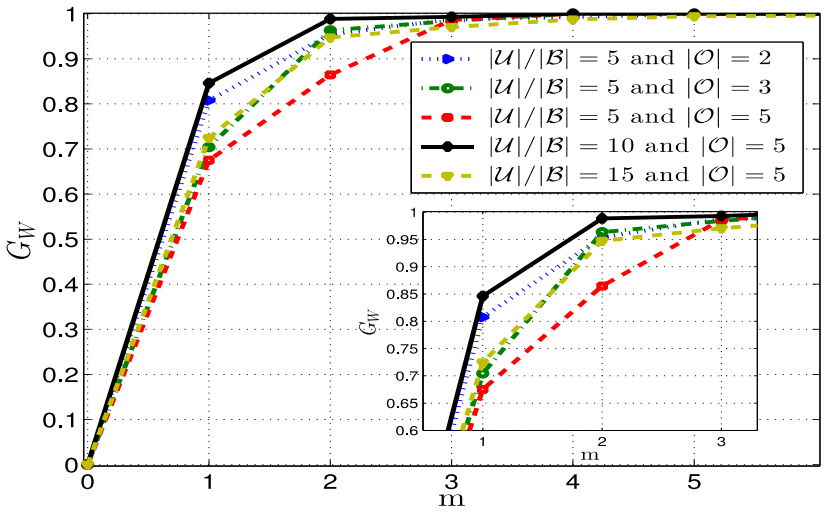


Figure 7-13: Normalised utility gain as a function of m

As shown from in Figure 7-13, the normalised utility gain obtained with *m* reassociations increase very sharply with GLLG. Furthermore, its complexity is very small as compared to MORA Non-linear solver, centralised algorithm with performance guarantees and GD, as shown from the results in Figure 7-14. The results confirm that Non-linear Solver and Centralised algorithms are impractical, especially taking into account that they have to be trigger every time the channel quality of a user changes. By contrast, the execution time for DG is very low and it's even lower for our GLLG approach.

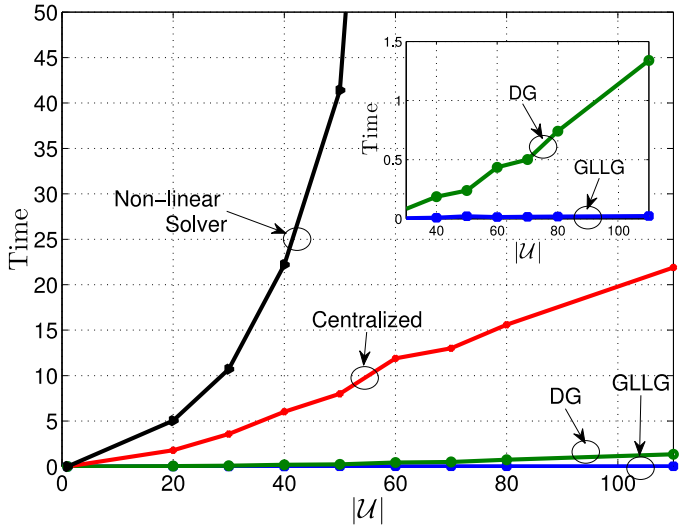


Figure 7-14: Computational complexity of GLLG and SoA algorithms

In terms of Network utility our approach performs very close to the benchmark given by a centralised algorithm and GD and it outperforms static slicing (SS) very substantially. The results are shown in Figure 7-15.

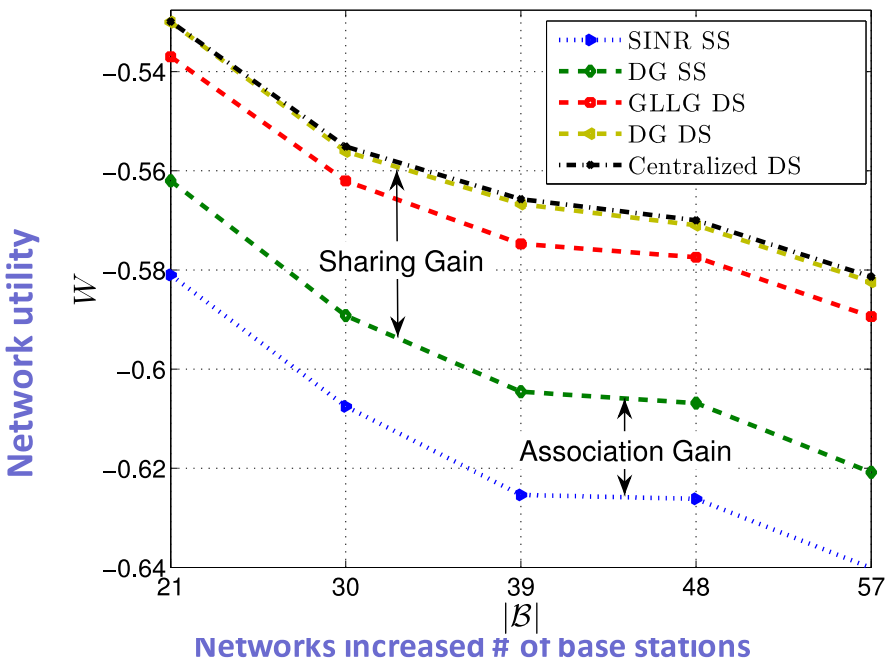


Figure 7-15: Utility gains for different approaches as a function of network size

We also obtain very substantially capacity saving that increase with the number of operators and with the density of base stations.

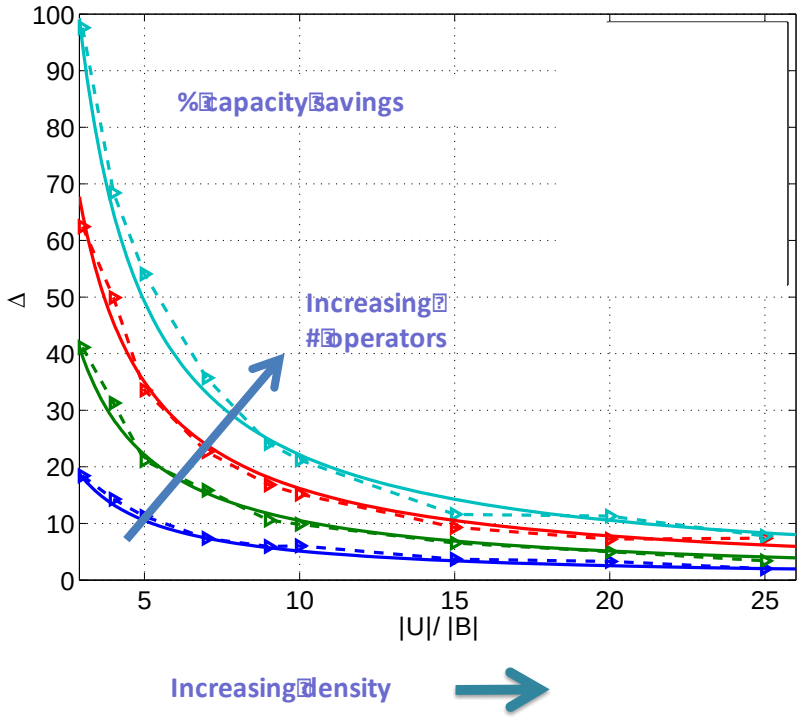


Figure 7-16: Capacity saving

To evaluate the gains from a user perspective, we compare the per-user throughput achieved by our approach against the static slicing (SS) with SINR-based user association (Baseline 1) and SS with enhanced user association (Baseline 2). We observe that our approach provides substantial

gains both in terms of the median values as well as the various percentiles. Similarly, to above the gains increase with the number of the operators.

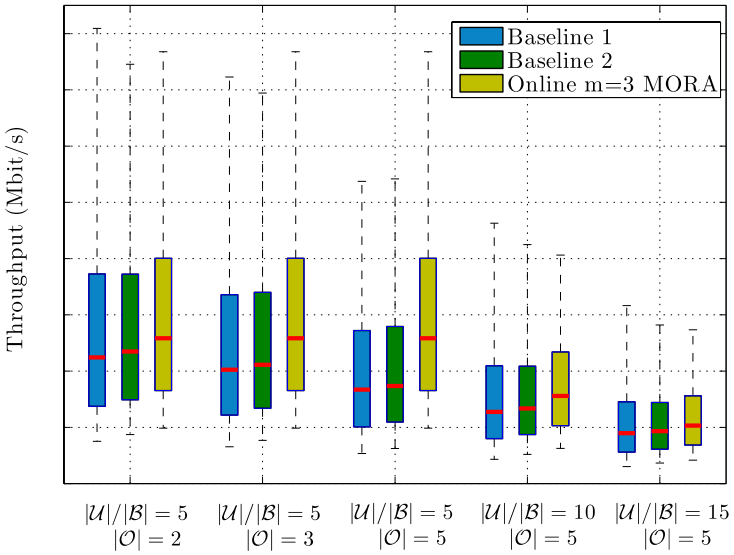


Figure 7-17: Improvement on the user throughput

We can conclude that dynamic resource sharing among tenants can be very beneficial. We have proposed a novel criterion that shares resources fairly among tenants taking into account their share, and the resources of each tenant fairly among its users. We have then devised a practical algorithm with limited complexity and overhead and which performance are very close to the optimal one.

7.6. Multi-RAT Integration

7.6.1. State of the art

Currently, mmWave technology like LMDS (Local Multipoint Distribution System) is either used for transmission of broadband data from one central point to homes and businesses as a line-of-sight transmission system (point to multi-point) or as a point to point transmission system, e.g. WiGig or 802.11ad within the 60 GHz-band [SDM+12] .

The mmMAGIC project is working on mmWave network deployment and integration topics, which are carried out in two directions: standalone and non-standalone operation of the mm-wave RANs. In the standalone scenario, the deployment of APs and relations between them will be optimised to mitigate coverage and mobility issues related to the mm-wave propagation. In contrast, the non-standalone operation scenario will utilise joint deployment of mm-wave nodes with nodes operating on lower frequencies (in the form of multi-RAT APs or neighbouring APs of different RATs).

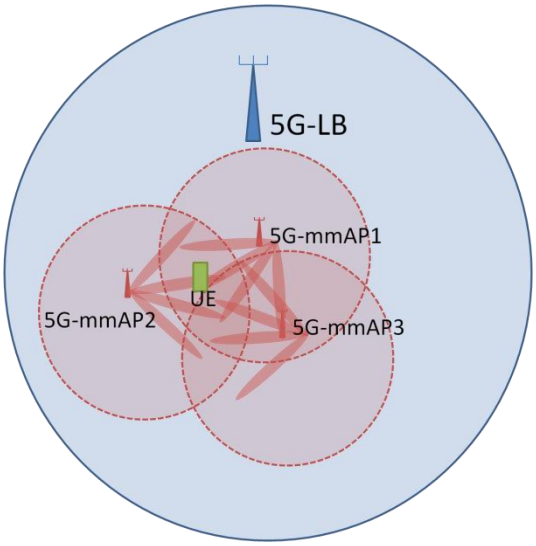


Figure 7-18: Deployment scenario: 5G eNodeB and mmWave APs

7.6.2. Detection of mmWave radio cells

Within LTE, each UE can detect radio cells which can take over the role of a serving eNodeB. The detection is based on common pilots and will lead to an event based reporting (e.g. A3). A mmWave systems will transmit UE related data by beamforming. Using uncoded pilots for mmWave detection will drastically reduce the coverage of the mmWave access point. We assume that precoded pilots at specific time slots will be provided by the mmWave access point, and all UEs within the vicinity of the mmTX point should be informed about these precoded pilots.

As a result, new requirements are imposed, which are listed in the following:

1. Specification of initial access schemes for mm-wave systems supporting high gain beamforming antenna configurations, based on beam search algorithms using precoded pilots and UE random access procedure.
2. Efficient tracking of beam direction (beam switching) in case of varying channel conditions or blocking effects
3. Measurement of target mmWave access points in case of UE mobility.
4. For these requirements a low band 5G node (5G-LB) should control the UE, i.e. new RRC functionality or new RRC protocol elements
5. Definition of a mmWave cluster (possible mmWave access points for each UE) by the 5G node and configuration of this cluster towards the UE.
6. Definition of time, frequency of precoded pilots: info to mmWave APs and UEs
7. Update of mmWave clusters in case of UE mobility.
8. UE measurement configuration and UE measurement evaluation for mmWave.

7.6.3. mmWave data handling

Within LTE dual connectivity (option 3C, split within PDCP layer) all data is processed and stored within the master NodeB (macro cell). For mmWave this would require large storage capacity within the 5G control node and many high-speed links to the mmWave APs.

We propose that within 5G the data storage and forwarding functionality should be revisited, e.g. a PDCP storage functionality should be defined, which forwards the data to the current serving mmWave AP, but which is controlled by the 5G control node/functionality as UE measurements and mmWave AP selection are processed within the 5G control node. This reduces also the burden of data transfer during handover between mmAPs and 5G control nodes.

Details of this proposal can be found in Part I of this report and in [AGA+16].

7.7. Data-layer and Control-layer Design for Multi-Connectivity

7.7.1. State of the art

The concept of multi-connectivity is standardised in LTE, under a technology known as Dual Connectivity. The term Dual Connectivity stems from the fact that the UE is connected to *two* eNodeBs at the same time. Next, we distinguish between the following three categories regarding the state-of-the-art pertaining to heterogeneous multi-connectivity approaches: a) Dual Connectivity features already included in LTE standards; b) Dual Connectivity features under discussion in LTE; c) Features under discussion in other research projects.

Standardised Dual Connectivity features in LTE [36.842]: Such features pertain to use case scenarios involving connection to a macro cell and a small (pico) cell.

Based on the *data-layer* architecture, the following options were recommended.

i. *Option 1A*: This option involves standalone Master (MeNodeB) and Slave eNodeBs (SeNodeB), in the sense that both MeNodeB and SeNodeB are connected with the core network (CN) via separate S1 interfaces. It was qualified in LTE since it yields a simple implementation with low backhaul requirements.

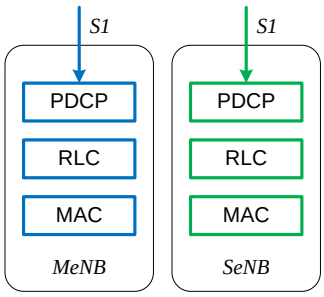


Figure 7-19: Option 1A in LTE

ii. *Option 3C*: This option involves SeNodeB which contains only the RLC and MAC layers. The S1 interface terminates at the MeNodeB, and involves splitting the bearer between the MeNodeB and SeNodeB. It was qualified in LTE since it yields the highest throughput among the candidate options. In fact, option 3C is the recommended dual connectivity solution in scenarios with high backhaul capacity.

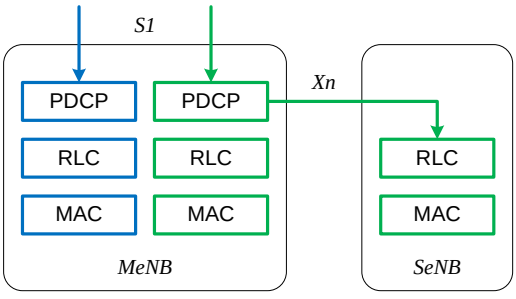


Figure 7-20: Option 3C in LTE

Based on the *control-layer* architecture, the following options were recommended

i. Option C1: Radio Resource Control (RRC) messages sent to UE are generated exclusively by the MeNodeB. Correspondingly, the UE replies control messages to MeNodeB.

ii. Option C2: Both MeNodeB and SeNodeB generate control messages. The UE replies accordingly to both MeNodeB and SeNodeB. How and whether to distinguish source and destination RRC entity was left for further study.

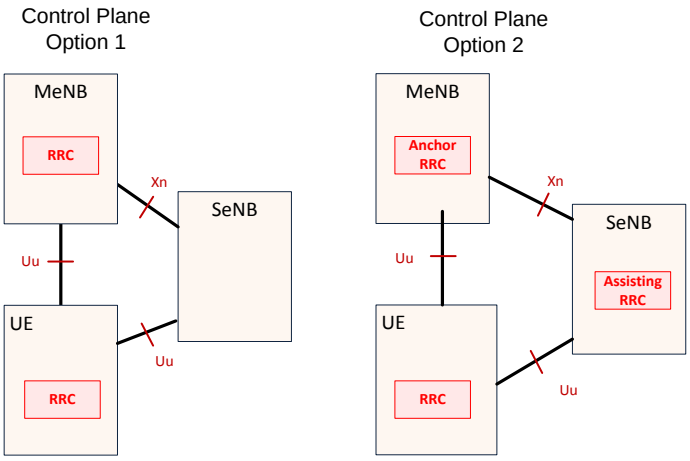


Figure 7-21: Options C1 and C2

Features under discussion in LTE: *Dual Connectivity for optimised handover*: This option was discussed in LTE standards but still no agreement has been made. It involves a SeNodeB before executing the handover, in order to smoothen the handover process and minimise the interruption caused. That is, the UE maintains its connection with a SeNodeB together with that of source MeNodeB until the handover to the target MeNodeB is completed. (Possible enhancement in LTE Release 13).

Features under discussion in research [METII-R21]:

i. *Coexistence with legacy technology*: Within the METIS-II framework, investigation of approaches involving connection with two different radio access technologies (inter-RAT) is being carried out. Particular focus is put on the case where the 5G RAT and the existing LTE technology are employed, i.e., when the UE is simultaneously connected to the existing LTE and 5G cells. Such approaches involve centralised/decentralised inter-RAT radio resource management. Moreover, the coexistence of 5G and legacy LTE network components is under study.

ii. *Support of fast activation of multi-connectivity*: In LTE, the use of multi-connectivity is limited by signalling procedures between MeNodeB and UE and MeNodeB and SeNodeB respectively. As the expected data rates are anticipated to be increased significantly in 5G, a fast activation of multi-connectivity is considered.

7.7.2. Towards Supporting 5G Multi-Connectivity

7.7.3. Overview of LTE Radio Access Network (RAN) Architecture

The main points pertaining to the RAN architecture considered in LTE standards are described in the following. First, the LTE RAN architecture consists of evolved Node Bs (eNodeBs), which, from the User Equipment (UE) perspective, comprise the data layer and control layer network termination points. That is, the network functions of the core network (known as Evolved Packet Core (EPC) Network) are not visible to the UEs.

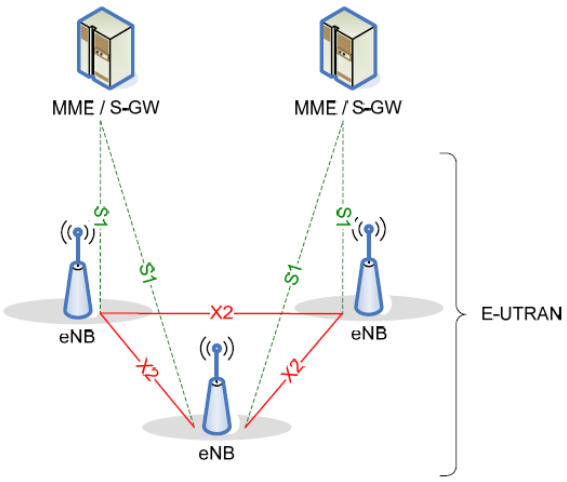


Figure 7-22: The LTE RAN architecture

Further,the eNodeBs may be interconnected with one another via a special inter-node interface, known as the X2 interface. The X2 interface is set such that inter-eNodeB functions are supported, including inter-cell interference coordination (ICIC), enhanced mobility, as well as several functions which belong to the broad family of Self Organised Network (SON) functions.

The eNodeBs support the full protocol stack, consisting of the following protocol layers: Packet Data Convergence Protocol (PDCP); Radio Link Control (RLC); Medium Access Control (MAC); Physical Layer (PHY). On top of these protocol layers the Radio Resource Control (RRC) layer is placed; its functionality is related exclusively to control layer messages.

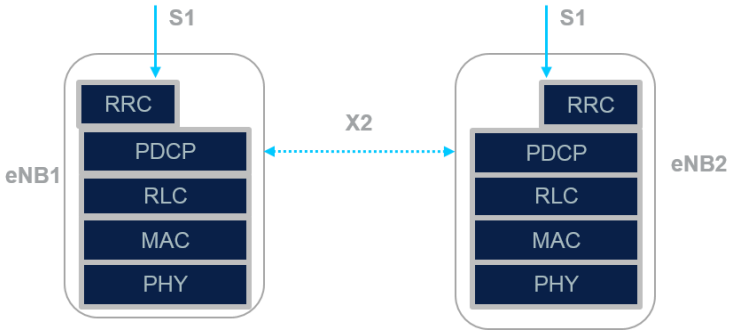


Figure 7-23: The LTE eNodeB protocol stack

The interface between the eNodeBs and the EPC network is defined as the S1 interface. Overall, it can be argued that *the LTE RAN architecture is fully decentralised.* That is, the eNodeBs are *standalone entities* which, albeit interconnected, they are independently connected to the core network.

7.7.4. Potential Shortcomings of LTE RAN Architecture for Multi-Connectivity Applications

Overview of multi-connectivity and its standardisation in LTE: Multi-connectivity involves the simultaneous connection of the UE to at least two eNodeBs. It particularly applies to Heterogeneous Network (HetNet) scenarios, associated with different layers of network coverage. In the most common scenario, multi-connectivity involves the combined connection of a UE to a wide-area macro cell and one or more small-cell(s). Multi-connectivity is documented in LTE standards

in Release 12: Referred to as dual-connectivity (DC), it involves the case where the UE is connected to two eNodeBs. Its operation is mainly related to increasing throughput via dual bearer connection and/or bearer split.

Why is LTE RAN architecture not suitable for multi-connectivity? From our viewpoint, the LTE architecture is not suitable for supporting the multi-connectivity requirements mainly for the following two reasons: a) Increased signalling overhead and b) support of ultra-reliable applications. These points are elaborated in the following.

Signalling overhead due to mobility: 5G network topologies are anticipated to deploy several clusters of 5G small cells, the coverage area of which overlapping with that of a (either 5G or legacy LTE) macro cell. Although not clearly defined yet, the number of 5G small cells within one cluster is expected to be large (i.e., some 10s of small cells per cluster), owing to their limited coverage area. Networks with such topology are known as Ultra Dense Networks (UDN).

The limited coverage area of small cells is associated with an increased occurrence of mobility events (such as handovers, cell measurements, etc), particularly for fast moving UEs. The frequent occurrence of mobility events entails a huge signalling overhead to the RAN, involving a set of control signals associated with handover commands are exchanged between eNodeBs. Additionally, the current RAN architecture allows that the frequent mobility events affect the core network as well. This is because each time a handover is triggered by the RAN the core network has to switch the transmission path accordingly. This raises the concern that the current LTE architecture is not suitable for supporting Multi-Connectivity in HetNets, unless a substantial signalling overhead is tolerated, involving both its RAN and core network.

Support of Ultra-high reliability: Dual connectivity in LTE standards focuses on increasing the throughput by establishing dual bearer connection to the UE. In some cases, bearer split is also supported, in the sense that the UE is able to split its bearer connection to two eNodeBs, aggregating thus its throughput. Nonetheless, in LTE standards no care was taken for addressing ultra-reliability scenarios (i.e., scenarios where high reliability is more critical than high throughput), since there was no such requirement in LTE. Ultra-reliable applications involve the *duplication* of one or more bearers across multiple eNodeBs, exploiting thus the concept of diversity. On the basis of the LTE RAN architecture, a bearer duplication would involve new features which would also increase the complexity of the corresponding deployment.

7.7.5. Proposed Architecture for Multi-Connectivity

In view of the above, it is evident that a novel architecture approach towards the support of multi-connectivity in 5G systems is needed. To this end, the proposed architecture involves the use of a RAN cloud, where the RRC (control) and the PDCP layer will be located. The remaining protocol stacks will remain on the eNodeB site, as shown in Figure 7-24.

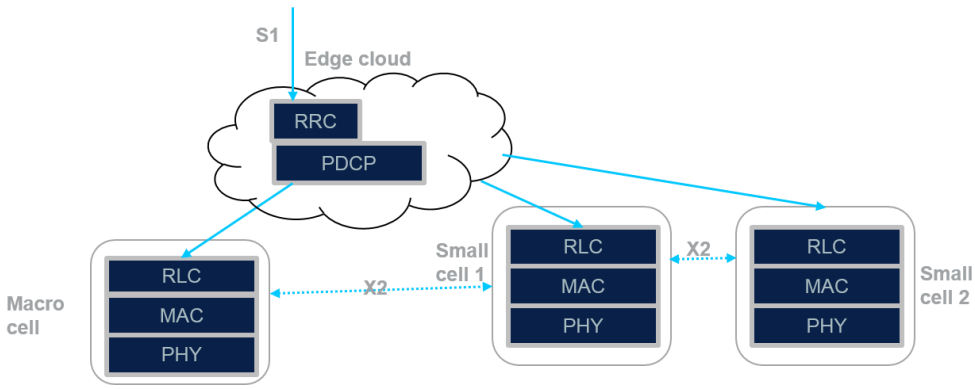


Figure 7-24: The proposed RAN architecture

With respect to the multi-connectivity-related shortcomings of the LTE architecture, the proposed architecture offers the following advantages:

- The frequent mobility between the small cells is hidden to the core network. This is because from the core network's perspective no path switch occurs each time a handover between two small cells takes place. In addition, the RRC layer where such the mobility of the UE is anchored remains the same. This results in a considerably lower signalling overhead.
- Data duplication across cells is facilitated: The PDCP layer in the RAN cloud would be responsible for duplicating the data across multiple cells. Such feature can be more easily supported with the introduction of the network cloud, resulting in much lower burden compared to duplication from the core network.

7.7.6. List of Relevant Network Functions

The network functions, which are related to the proposed architecture for multi-connectivity, are listed in the following Table 7-1. They are categorised into functions already existing in LTE but need to be adjusted appropriately, and new functions which need to be introduced in 5G.

Table 7-1: List of relevant functions

	Layer	LTE functions to be modified	New functions for 5G
	RAT Selection		
	SON		
	Network Mngmt		
	Mobility Conn. Mngmt		
RAN	RRC		
	PDCP	-Data Transfer -Routing and Reordering	-Mapping between service flow and radio-bearer service for enhanced QoS support and in-service-flow differentiation -Routing and flow control for enhanced RAN level multi-connectivity with possible PDCP level radio bearer split and cloud-RAN support -Further anchoring functions for flexible, on-demand data-layer enhancements, including security and mobility on demand, data-layer/control-layer separation and cloud-RAN support
	RLC	-Buffering/transferring of PDCP PDUs -Reordering of RLC PDUs	

		-Duplicate detection of RLC PDUs -Reassembly of RLC SDUs -Re-segmentation of RLC Data PDUs	
	MAC	-Link adaptation	-Scheduling info exchange -Common priority handling -1x uplink coordination -UE radio network identities
	PHY		

7.8. Flexible 5G service-flow (SF) with in-SF QoS differentiation and multi-connectivity

7.8.1. State of the art

METIS II R2.1 for RAN design guidelines [METII-R21] serves as a good summary for the state of the art related to 5G RAN design and development. Further to [METII-R21], the following QoS/QoE and multi-connectivity aspects are considered.

QoS/QoE aspect: In LTE, packet radio services are provided based on a bearer service model in which multiple concurrent radio access bearers (RAB) need to be set up and used for c-layer and d-layer connection [23.401]. Each RAB consists of a S1 bearer and a radio bearer (RB) with 1:1 mapping relationship. RAB setup and related QoS control is primarily based on policy enforcement at the core network (CN) side which is initiated and controlled by MME with some application agnostic QoS enforcement on RB at the air interface.

Figure 7-25, i.e., Figure 4.7.2.2-1 of [23.401], illustrates bearer service model and QoS concept in LTE.

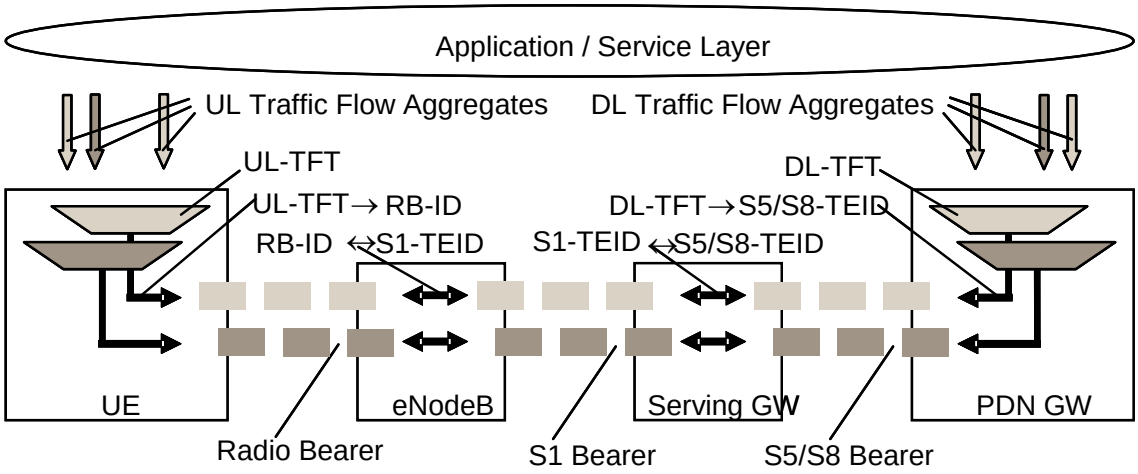


Figure 7-25: Illustration of LTE bearer service model and QoS concept

The bearer specific QoS parameters are QCI (QoS Class Indicator), ARP (Allocation and Retention Priority) and data rate specific parameters including GBR (Guaranteed Bit Rate) and MBR (Maximum Bit Rate) for individual GBR bearers.

The ARP is mainly used in the case of congestion for deciding if the bearer establishment or modification requests could be accepted. The ARP is also used by the eNodeB to decide which bearers are dropped in sudden resource shortages. The ARP contains the pre-emption vulnerability and pre-emption capability fields that define if the resources of the bearer could be allocated to the higher priority bearer.

QCI is a pointer to record of node specific performance parameters, which are Resource Type, Priority, Packet Delay Budget and Packet Error Rate.

The current LTE bearer service model and E-UTRAN as such is not flexible and granular enough to respond to 5G service requirements and expectations including supports of challenging new 5G services and enhanced experiences with massive-grown OTT and Internet traffic in both application diversity and volume [MET-D11], [DNS], [MDU]. It is stated in [TAT] that "virtually every network-capable application in existence today relies on TCP/IP to function properly".

Hence, it is desirable to have 5G RAN design also optimised for supports of TCP/IP traffic and OTT applications on per UE per application level.

Multi-connectivity aspect: In LTE either intra-LTE dual connectivity or RAN level integrated inter-RAT LTE-WiFi dual connectivity is introduced, primarily for boosting data rate for UE utilizing small-cell carrier while active connection and mobility management is maintained and controlled by a serving macro cell [36.300].

5G networks and services with different air-interfaces and ultra-low latency and high reliability requirements may explore multi-connectivity capabilities in a more flexible and effective way in order to provide the UE with better data rate as well as to fulfil latency-reliability requirements of new 5G services in both c-layer and d-layer. Furthermore, as tight interworking between 5G and LTE is considered as a requirement for 5G, 5G multi-connectivity involving LTE needs to be considered which may have certain implications on both LTE and 5G networks.

Expected new network functions: The same list of expected new network functions as provided in Section 7.7.2 is applied, in addition to other relevant network functions and function blocks described in previous chapters.

Further enhancements on MAC functions and services for MAC level MC and QoS supports may be expected, as 5G RAN and MAC in particular may have to support radio access modes and transmissions with much broader dynamic range in terms of bit rate, latency and reliability requirements, as compared to that of LTE.

7.8.2. RAN support for advanced QoS/QoE control

7.8.3. Facilitating in-bearer QoS differentiation

First, let us clarify some terms and make some assumptions, as follows:

The term bearer is inherited from and therefore referred to the bearer concept of LTE, including radio bearer (RB) in E-UTRAN.

The term 5G service flow (5G SF) is referred to a logical connection in 5G UP between an active UE and a serving GW in d-layer, denoted as uGW, which consists of a radio access link or connection between the UE and a serving RAN and a transport network path between the serving RAN and the serving uGW. 5G SF as such may be broader or more flexible than EPC bearer of LTE, in term of tunneling and mapping between EPC bearer and radio bearer as well as logical service flow resolution inside a SF.

UE in connected and active state may have one or more SF established toward one or more serving uGW. The establishment of a SF may involve control from a serving GW in c-layer, denoted as cGW which may be same as or different from uGW.

5G SF may be local, i.e., not actually routed via the serving uGW but an optional local gateway functionality closer to RAN (in between uGW and RAN). The term 5G elementary flow (5G eF) is referred to the finer or lower level of the SF resolution which is meaningful to 5G QoS/QoE control framework, or, that is, 5G allows for filtering, monitoring and controlling till individual eSF level. The term 5G sub-flow (5G sF) is referred to a group or subset of eFs within a SF which have some common predefined attribute(s). QoS control can be in different levels such as RAN, CN and E2E. RAN level radio-link specific control may be independent from CN level backhauling connection specific control to some certain extent.

Let us provide some examples of SF, sF and eF in practice. Mary is using her 5G smart phone UE for working, socializing and getting multimedia contents from the Internet. Her UE may have two SFs established, one toward her private enterprise service network and another toward public service network. In her first SF toward the private enterprise service network, there is a single sF configured for all background active office applications such as emails and so forth. Note that this is just for a simplified illustration without considerations of UL/DL or two directional communications needs. In her other SF, there are currently two SFs configured, one for an ongoing real-time multimedia call and another for all remote Internet access services. Let us assume that Mary is currently active on, e.g., Facebook, uploading a new photo album and a set of video clips while chatting with a few others. In parallel, she is also checking some online newspapers and downloading some film and there are full of ads on her browser screen as well. These mean there can be tens or even hundreds of eFs within the second sF of the second SF.

It is highly desirable that 5G QoS/QoE framework may enhance end-user experience for OTT and Internet applications: granular QoS/QoE enforcement on per user/application level. RAN supports for tightly coupled DL/UL QoS/QoE enforcement actions on dynamic sub-flow level operations and services need to be considered. The main practical questions are which specific logical levels of service flows should be considered and how to handle the QoS differentiation with possible multi-connectivity based on pre-defined logical levels flexibly and effectively, including specific CP-UP interactions and signalling procedures in UP and CP between UE and serving network.

In 5G NORMA, we propose that the CP configures and controls UP on 5G SF and sF level, meaning that contexts of SF and sF(s) as well as L:M mapping between SF and sF(s) on RAN and CN level are established and maintained by CP. Thus, from RAN point of view, sF(s) is similar to RB(s) of E-UTRAN which is (are) configured to the serving RAN for RAN level transmission and QoS control (1:1 mapping on MAC LCID – Logical Channel Identifier – and corresponding RLC entity). The in-bearer or in-SF/in-sF QoS differentiation is on the individual eF level which is kept within UP and primarily managed and controlled by PDCP for RAN level and some UP master QoS handling entity in CN such as uGW for both CN and RAN. That is, CP (RRC or NAS) will not maintain the contexts of individual eFs. This way allows for flexible and scalable handling of UP with possible in-bearer or in-SF QoS differentiation which has a clear enough CP/UP resolution aiming for not causing much additional complexity or overhead to CP. This also allows for flexible implementation, ranging from e.g. a simple plain option of having 1 SF mapped on 1 sF and no eF per UE to an advanced option of having L SFs mapped on M sFs with N eFs getting certain QoS differentiation treatments per UE. Note that the LTE equivalent option is L SFs 1:1 mapped on L sFs and ignoring all eFs.

CP (RRC or NAS) may assist UP in signalling eF related control information between the serving RAN or CN and the UE, as triggered and requested by UP (PDCP in RAN or UP master QoS handling entity in CN such as uGW). In this option, detailed information on the eF of interest is provided to CP (RRC or NAS) by UP (PDCP or uGW). The CP assistance may be done during sF establishment during which the related control information of expected eFs may be signalled. Or the CP assistance may be done when an eF of requiring certain QoS differentiation treatment

is identified by UP. In addition, new UP control signalling procedures on eFs may be introduced to PDCP between RAN and UE using, e.g., different PDCP C-PDUs or PDCP PDU header fields. This option allows for the eF level control to be kept within UP and therefore not causing any notable impact on CP.

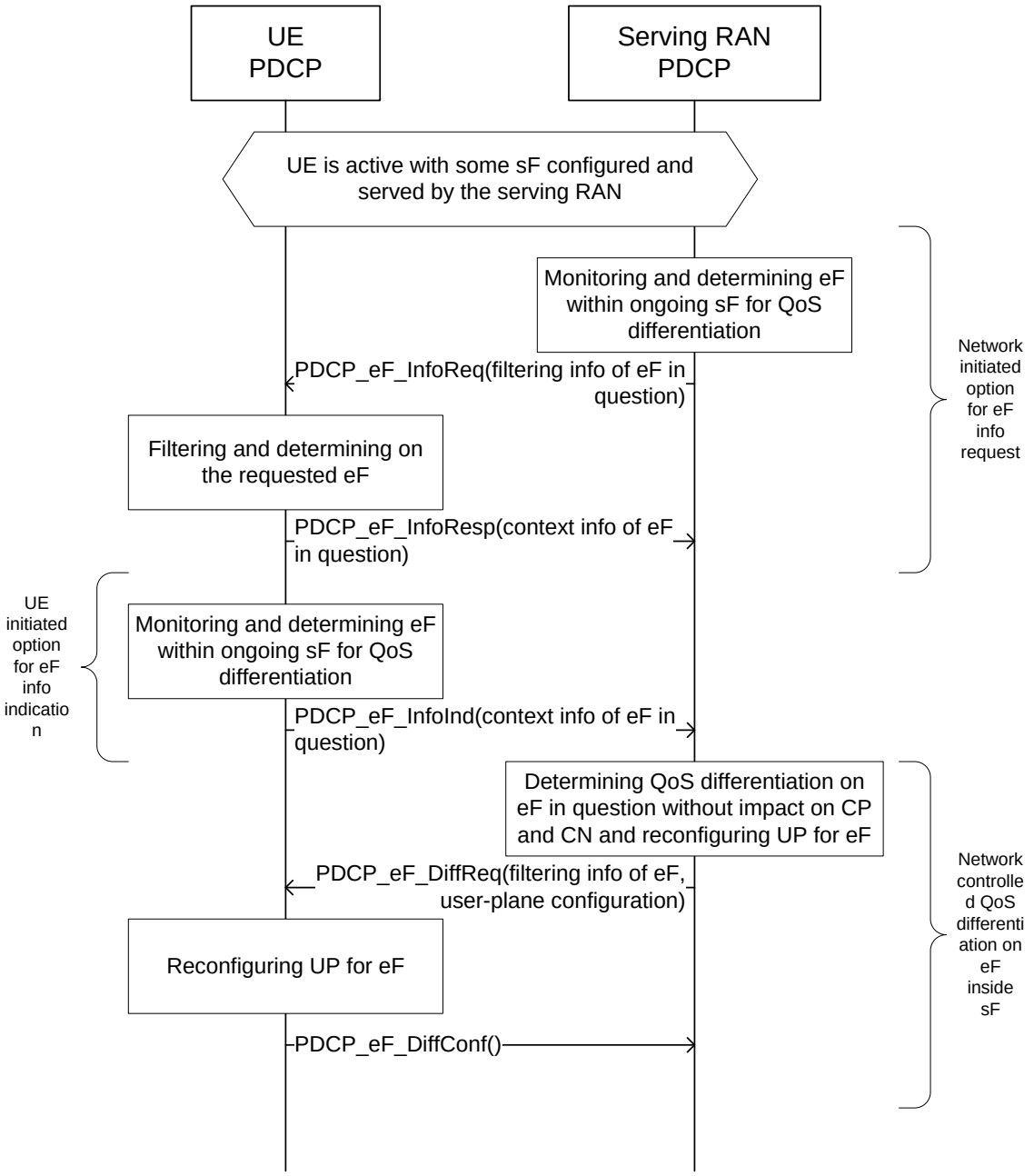


Figure 7-26: Illustration of d-layer PDCP signalling procedures

The network side, based on received eF information, may determine the actual QoS differentiation treatment on the eF and configure that to the serving RAN and UE in order to carry out the determined QoS differentiation treatment.

In 5G NORMA, we propose two instances of in-bearer QoS differentiation, i.e., QoS differentiation treatments requiring CP-UP and RAN-CN interactions, and QoS differentiation treatments in RAN level and not requiring CP-UP and RAN-CN interactions. QoS differentiation treatments requiring CP-UP and RAN-CN interactions aims for those treatments which involve both RAN

and CN for E2E QoS/QoE control operation and allow for possible UE negotiations as well as charging impacts on-the-fly. By contrast, the second case, QoS differentiation treatments in RAN level and not requiring CP-UP and RAN-CN interactions, requires UP-decided remapping or re-routing of the eF from the current corresponding sF onto the other established sF of the UE which has more suitable configured QoS attributes and constraints. PDCP may use its own signaling procedure or request RRC to configure this remapping or rerouting between the serving RAN and UE. Note that in the multi-connectivity contexts, the UE may be served by more than one AP and therefore remapping or rerouting of the determined eF between 2 existing sFs may involve a change of AP as well.

Figure 7-26 illustrates a PDCP level control signalling procedure within UP for facilitating some in-SF QoS differentiation. Figure 7-27 illustrates a more extended control signalling procedure involving both CP and UP for facilitating some in-SF QoS differentiation.

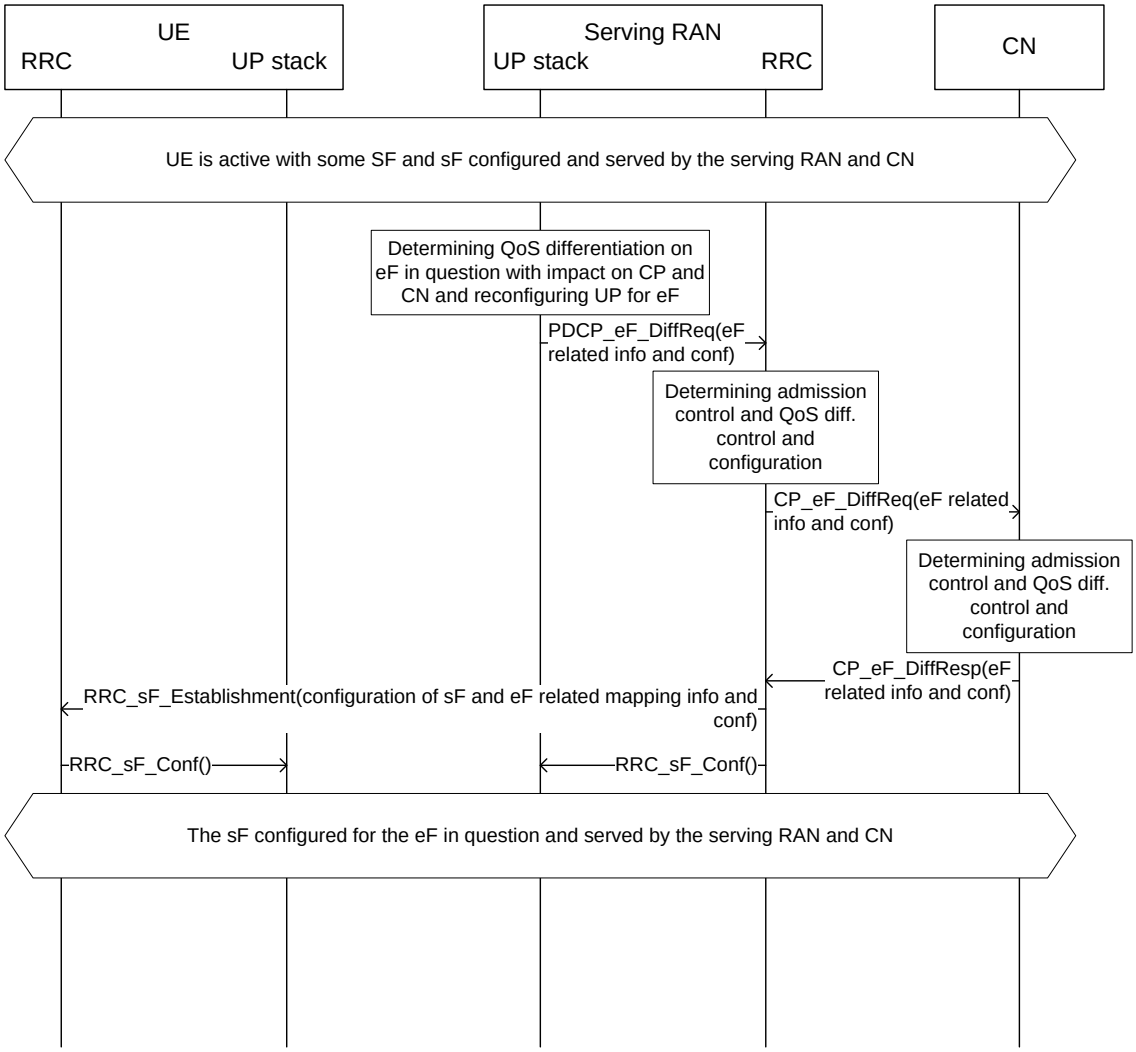


Figure 7-27: RAN triggered d-layer and c-layer interactions and procedures

SDMC Functions:

SDMC NAS Control and QoS Control may interact with RRC User to facilitate this innovation, as illustrated in the extended procedure of Figure 7-27.

7.8.4. D-layer enhancement for optimised QoS supports

More than 90% of today's Internet traffic, including most of popular social networking and multimedia sharing applications, is using TCP/IP. Hence, enhancing TCP performance in general and QoS/QoE of TCP/IP based applications and services over 5G network access is an important design target for 5G networks.

However, a TCP connection is a bidirectional end-to-end transport-layer connection between the data source and the data sink with known properties or features such as TCP version, slow-start and robust flow control but is very sensitive to packet errors, latency, or round trip time (RTT). TCP performance is particularly sensitive to latency associated with successful transmission of the first packets of the TCP connection associated with the connection establishment procedure (also called as TCP three-way handshake: SYN, SYN-ACK, ACK) which often does not carry actual payload and therefore has small size in packet length (can be in extent of TCP header).

A RAN d-layer protocol stack and PDCP in particular, as assumed to be somewhat TCP/IP and application aware, is able to monitor and filter out individual TCP/IP packets and sub-flows as well as carry context information of upper layers passed down to PDCP by corresponding upper layers. PDCP at the UE side is the first or earliest possible RAN-level entity which may be configured to monitor and filter out individual application packets and service flows originated from UE for the uplink (UL) traffic.

Hence, d-layer enhancement to speed up transmissions of targeted "special" d-layer packets such as first packets of TCP connection or C-PDUs of UP protocols on RAN level is proposed, as illustrated in Figure 7-28.

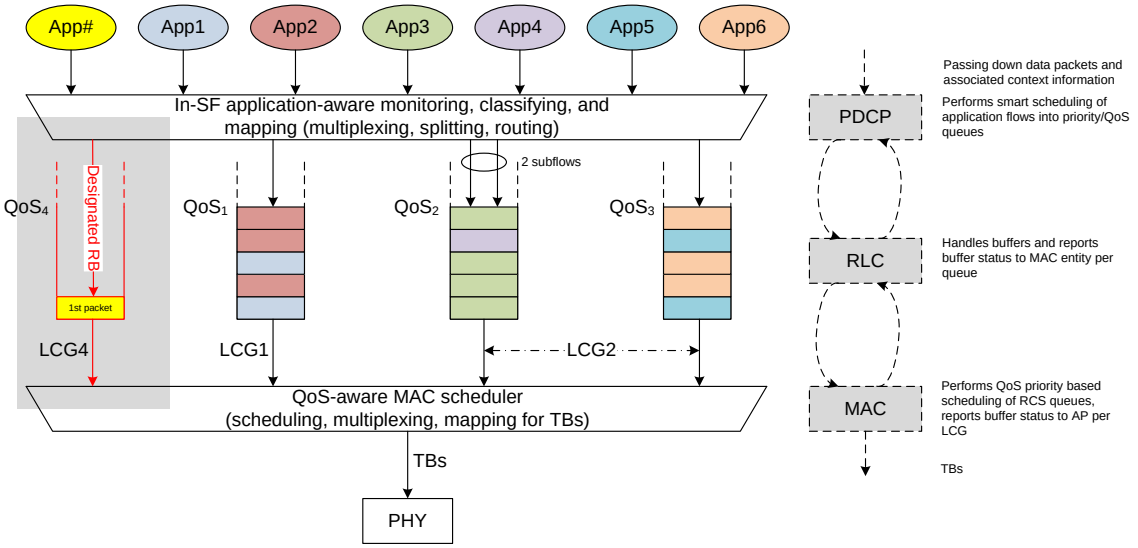


Figure 7-28: Illustration of UL with the designated RB marked in red

The proposed method is based on the idea of having a designated UL L2 priority queue (PQ) or radio bearer (RB) or logical channel (LC) set up and used for transmissions of all targeted UP packets by targeted UEs in UL. The same analogy may be applied for the DL as well. That is, the active UE is configured by the serving network to set up and use a designated RB (PQ or LC) in UL for transmitting some specific UP packets such as first packets of TCP connections (established and used for corresponding TCP/IP based applications of the UE) or Control-PDUs of PDCP to the serving RAN. The designated RB is associated with highest possible priority as compared to that of other RB used for transmissions of other or further UP data (see red queue in Figure 7-28).

L2 PDU format (header or add-on control element of PDCP PDU for instance) as specified for the proposed RB may be different, compared to that of other RB types, in order to carry specified context information associated with the corresponding new TCP flow of the individual first packet sent in the payload of the same L2 PDU. The specified context information of the corresponding TCP flow may include, e.g., application related type or QoS profile ID mapped or set by UE based on information passed down from application layer and/or preconfigured mapping rules/policies, source-sink direction indication, ranges of expected data volume or connection lifetime, and so forth. The specified context information may implicitly or explicitly indicate initial mapping of the corresponding TCP flow on another established RB for subsequent TCP packets, e.g., it may specify an already configured RB/queue number or the said QoS profile which maps to an existing RB or triggers configuration of a new one. This is considered as UE assistance information which the serving network may use when configuring the UE and handling the newly created service flow.

Using the designated RB, the serving RAN is not only able to receive and deliver those targeted packets more quickly and reliably, but also to get desirable application-aware assistance information from the UE more quickly and reliably when it's actually needed. This helps the network side to detect, make decision and control on the individual TCP flows with possible QoS differentiation in a fast, reliable yet simple manner. For instance, consider the PDCP at the serving RAN upon receiving a PDU sent in the designated RB from the UE peer. The PDU is carrying a first TCP packet and further context information associated with the corresponding new TCP flow of the first packet It may have sufficient initial knowledge about the new TCP flow such as identity, further application-aware context, initial RB mapping as well QoS profile characteristics. Hence, the serving RAN may decide and carry out any necessary reconfiguration, remapping or other treatment on the new TCP flow and the RB on which the new TCP flow is to be mapped and transmitted.

In case the UE is served in radio multi-connectivity, the serving network may decide to configure one designated priority RB for all the radio connections of the UE or configure separate ones per corresponding radio connections of the UE. The UE may be configured to decide and route those "special" packets to a certain priority RB.

For the DL, the same analogy may be applied. However, some of the associated application contexts in the aforementioned UE assistance information (such as expected data volume or session lifetime) of the network-initiated flows in DL may not be available at PDCP together with the first packets of new TCP flows. Instead, the network-side PDCP may include QoS control information in the header of PDCP PDUs specified on the designated RB in DL, such as priority setting or RB mapping instruction for the associated TCP flow in the UL.

7.8.5. C-layer for flexible radio multi-connectivity

5G systems may cover different spectrum ranges including below 6 GHz, cmWave and mmWave, which may end up with very different new radio interfaces (RIs) as the physical nature of the spectrum ranges is different. With multi-RIs support in 5G, it is important to have the possibility of supporting separate RAN level configuration of different RIs as the functionality and configuration parameters of different RI might be different. This will also allow the development of different RIs not dependent with each other in different timelines. Therefore, there is a need of designing RAN control layer architecture that can adaptively support multi-RIs and multi-connectivity in an efficient and flexible way. Furthermore, reliability, efficiency and robustness of RRC control connection may need to be further enhanced for supports of ultra-reliable and low latency communications (URLLC) in 5G.

In LTE dual connectivity (DC), secondary eNodeB (SeNodeB) owns its radio resources and is primarily responsible for radio resource management of its cell. However, the final RRC message is generated in master eNodeB (MeNodeB) and sent to UE over the radio link of the primary cell that MeNodeB manages. There may be some drawbacks with LTE DC approach, such as more

configuration delay, extra effort on controlling of re-configuration timing, more processing overhead for MeNodeB and signalling overhead on radio link of MeNodeB. Furthermore, the MeNodeB needs to be aware of the radio interface of the SeNodeB (e.g. decode measurement reports containing dual connectivity events). In the context of 5G with multiple RIs, more radio connectivity links and multi-tenancy support, there are additional issues to be considered, e.g.:

- The entity that hosts the master control functions may be more easily overloaded due to more potential radio legs that one UE can support.
- The air interface of the master may be overloaded due to the increased amount of reconfigurations for SeNodeB addition/removal/modification.
- In supports of flexible and dynamic routing at PDCP level of 5G for multi-connectivity, the back and forth transfer of signalling messages in case that final RRC message is generated by the master RRC entity as in LTE DC but actually transmitted on the radio link provided by a slave-RRC may introduce unnecessary configuration delay and signalling overhead on backhaul

One of the multi-connectivity scenarios is illustrated in Figure 7-29, in which it may be beneficial to have independent RRC for certain connectivity among others. For example, an UE is initially in LTE DC with eNodeB1 as SeNodeB (no RRC instance is added in eNodeB1 as all the RRC messages are transmitted via MeNodeB) and eNodeB3 as MeNodeB. When the UE enters into zone to add cluster-eNodeB2 as SeNodeB and an RRC instance is added so that autonomous mobility procedure and bearer configuration using cluster-eNodeB2 radio link rather than using eNodeB3 radio link can be managed by the RRC instance. The cluster-eNodeB does not refer to any particular concept, rather just a group of co-located small-cell eNodeBs deployed under macro coverage of MeNodeB.

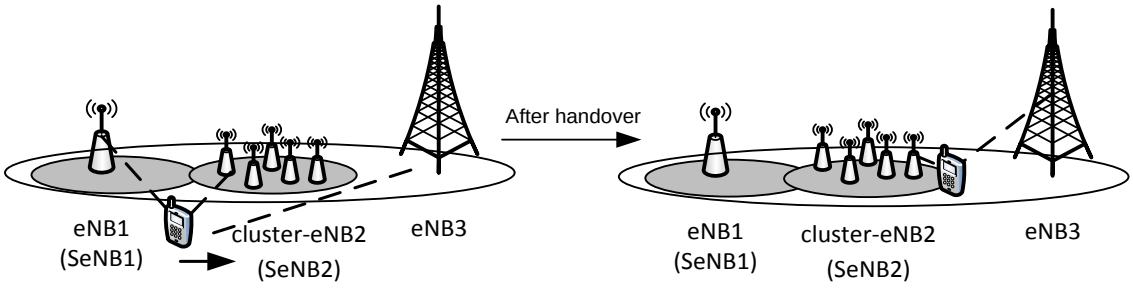


Figure 7-29: Example of multi-connectivity scenario that may prefer some independent RRC function at SeNodeB

Such a construction allows the following flexibility:

- Enables multi-connectivity by adding SeNodeB2 to already LTE-DC between eNodeB1 and eNodeB3
- UE mobility within SeNodeB2 is autonomous (i.e. MeNodeB resources are not involved neither does MeNodeB has to implement SeNodeB2 radio interface awareness)
- It is possible to reconfigure the radio leg of SeNodeB2 independently of the LTE-DC
- When UE fully leaves coverage area of eNodeB1, eNodeB2 could become SeNodeB with single reconfiguration message

Hence, 5G NORMA considers the coordinated RRC control structure in multi-radio and multi-connectivity scenario where the master-RRC, which may be located in either one of the 5G APs or RAN aggregator entity, will coordinate the RRC control among multiple radio legs. The slave-RRC, which is located in the remaining 5G APs, will have the possibility to manage of radio leg-specific control and procedure directly towards UE if it is established under the control of master-RRC.

The master RRC coordinates the RRC control among MC radio legs, including dynamic on-the-fly setup and release of a slave RRC which is then delegated for managing certain RRC control

procedures for the corresponding radio leg directly towards UE, as illustrated in the following figure.

The master-RRC may determine if a slave-RRC needs to be initiated or not during an additional radio leg establishment or radio leg reconfiguration for MC. Herein radio legs may be the same or different 5G RIs of the other radio legs. The determination may be based on multiple criteria such as

- UE capability in term of multi-RRC support;the location and processing load of physical network entities or nodes that master- and slave-RRC may be located;
- The multi-tenancy involvement of the cells that the radio legs are connected to;
- RI's characteristic of the radio leg;
- The channel condition or radio-link quality of each radio leg;
- The service flow characteristics;

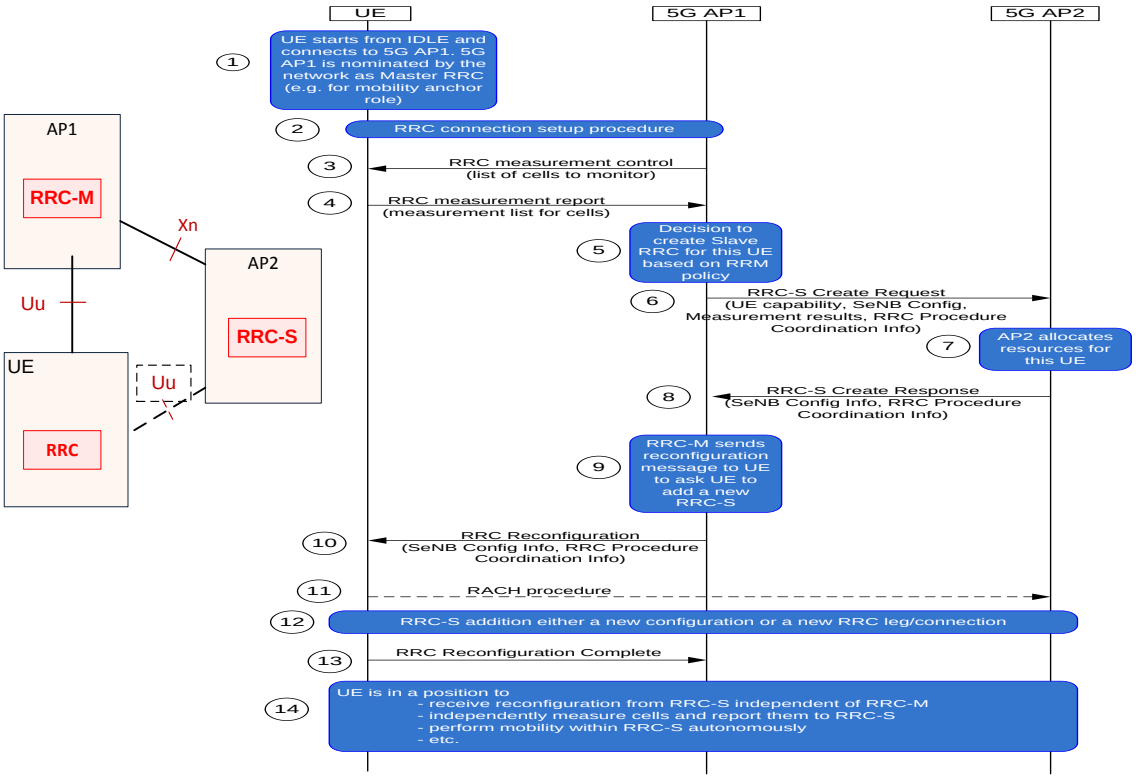


Figure 7-30: Illustration of master-slave RRC setup for controlling MC of UE

The master-RRC may coordinate the division or assignment of the RRC functions and procedures between master- and slave- RRC on the fly and on UE or service flow basis. The above listed criteria may be used here as well. For instance, slave-RRC may be configured to perform more RRC procedures if radio leg link quality is in good condition and/or the cell load is low. In another example, slave-RRC may be configured to perform less RRC procedures if RI of the multiple radio legs is the same and more dependent on each other.

The coordination may be performed in the network side over the interface (e.g. X2 kind of interface in LTE) between involved network entities. For instance, master-RRC may determine to initiate a RRC procedure toward UE via certain slave-RRC. The intended RRC message initiated from master RRC may include all detailed control information (full), part of detailed control information (partial), or no detailed control information (empty). Then slave RRC decides either to forward the RRC message as such, add further detailed control information, or fill in all detailed control information and send that to UE, corresponding to full, partial or empty indication in the RRC message. Furthermore, the full message option may also implies that corresponding RRC procedures should be terminated in master RRC; the partial option could be terminated in both

master RRC and slave RRC; and the empty option could be terminated in slave RRC. Master RRC can of course indicate to slave RRC explicitly if master RRC or slave- RRC or both should be the termination point of the corresponding RRC message or RRC procedure. To facilitate the coordination between master RRC and slave RRC, PDCP or X2 may be involved and such indication can also be used by PDCP to determine which radio leg the RRC message should be transmitted. For the full and partial RRC message initiated from master RRC, it may be signal to UE in parallel by both master RRC and slave RRC. In this case, UE may response to the earliest received or both for reliability.

In addition to the division or assignment of RRC procedures described above, master RRC may also indicate the configuration constraints that are under slave RRC control in order to avoid UE capability violation. The slave RRC may also request the update of the RRC functional division as well as the configuration constraints.

UE may be configured to facilitate the coordination of master RRC and slave RRC configuration. For example, UE may report the RRC configuration of each radio leg to the master RRC either periodically or as event triggered when detecting the conflict of different RRC configurations. Based on the UE report, master RRC may determine to reconfigure slave RRC of certain radio leg properly.

For RRC procedures that are under the control of slave RRC, UE may be configured to use the corresponding radio leg for UL RRC message transmission. For RRC procedures that are under the control of master RRC, UE may be configured by master RRC which UL RRC message can be transmitted via which radio leg or multiple radio legs. For instance, master RRC may configure UE to initiate RRC via slave RRC at the first place and switch to master RRC only for some specified unexpected events or once per a certain configured period.

7.8.6. MAC level multi-connectivity for ultra-dense 5G networks

Let us consider an ultra-dense network (UDN) which is deployed with high density of small cell access points (AP) over a certain hotspot service area in order to provide services including those with ultra-high reliability (virtually zero packet error rate) and ultra-low latency (as low as 1 ms radio latency) requirements. It is expected that in UDN a UE is, for most of the time, in an overlapping coverage of a number of local small cells and that the numbers as well as radio properties of local APs and UEs are rather comparable. In this regard, UDN environment is rather comparable to some proximity-based device-to-device (D2D) communication environment.

It is well known that broadcast-based connectionless communication is a simple and effective method for proximity-based group communications, as adopted in 3GPP Release 12/13 Proximity Services (ProSe) direct D2D communications for public safety (PS) uses for examples.

However, there is not much difference in terms of radio resource utilisation (RRU) between broadcast-based and unicast-based radio transmissions individually to each UE in UDN, unless sophisticated and often complicated beamforming is used. Furthermore, when it comes to providing radio MC in UDN, radio MC with a broadcast-based radio connection may actually consume less radio resources and have lower control overhead than that using multiple unicast-based radio connections.

In addition, the instantaneous nature as well as multi-receiver diversity of broadcast-based communication may be explored for enhancing or fulfilling challenging latency and reliability requirements.

By exploring advantages of broadcast communications in a proximity communication environment as of UDN, we propose a simple and effective radio MC scheme with broadcast-based UL and unicast-based DL transmissions for providing cellular access services with challenging la-

tency and reliability requirements. This aims to provide connection-oriented network access services which have challenging requirements such as ultra-low latency and ultra-high reliability for an active UE of interest in a serving 5G network.

First, UE is configured and controlled by a serving AP (5G-NB) to broadcast UL transmission of a TB which contains a MAC PDU to a group or cluster of targeted local small cell APs currently in proximity of the UE, as in proximity-based communication. The proximity-based UL transmission of the UE may be scheduled either by the UE autonomously – using resources from some preconfigured resource pool(s); or by the serving AP – either dynamically per each transmission or in a semi-persistent scheduling (SPS) fashion. There are also hybrid options, one of which, for an example, may allocate dedicated resources for the UE to indicate the scheduling assignment (SA) of the next scheduled UL transmission of which the resource allocation and transport format is determined by the UE autonomously using resources from some preconfigured resource pool(s). A UE may be configured to adapt the PDU format of targeted UL transmissions, depending on or adapted to which level, e.g., PHY TB, MAC PDU or MAC SDU, the packet is received and forwarded in the proposed radio MC scheme at the network side.

The approach further involves a dynamically coordinated and cooperative MC cluster (CCMCC) of local small cell APs (currently in proximity of the UE) configured to determine whether to receive proximity-based UL transmission of the UE and forward the received UL transmission of the UE towards a determined anchor of the radio MC. The CCMCC is therefore specific to the UE, also considered as UE-centric in the UE-centric networking paradigm. The MC anchor may be: one of the APs in the current CCMCC which can be the same or different from the current serving AP for the DL; or a MC controller node (MCN) which configures and controls the dynamic CCMCC, as illustrated in Figure 7-31. The destination address of the targeted MC anchor is determined by a forwarding AP in CCMCC based on either UE specific control information preconfigured to CCMCC using a network signalling procedure or packet specific control information such as destination address included in the header of received UL MAC PDU from the UE. The forwarding AP is configured to provide necessary labelling of the received packet to be forwarded to the targeted MC anchor, depending on or adapted to which level, e.g., PHY TB, MAC PDU or MAC SDU, the received packet is forwarded.

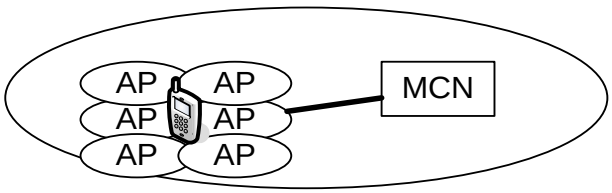


Figure 7-31: Illustration of UDN with MCN controlling CCMCC

Figure 7-31 illustrates a CCMCC, assumed to be configured and controlled by a MCN in serving the UE. Let us consider the case that MCN is also the MC anchor of the UE in UL and therefore receiving individual UL packets of the UE forwarded by one or more APs of CCMCC in serving the UL MC of the UE. MCN may dynamically track and configure CCMCC for the UE based on, e.g., knowledge about neighbour cell relation of the small- cell AP which is serving the UE in DL or a UE report of the discovered small cells or an AP report of discovering the targeted UE. CCMCC is considered as "liquid" cell cluster adapted to mobility of the UE in UDN (UE-centric mobility).

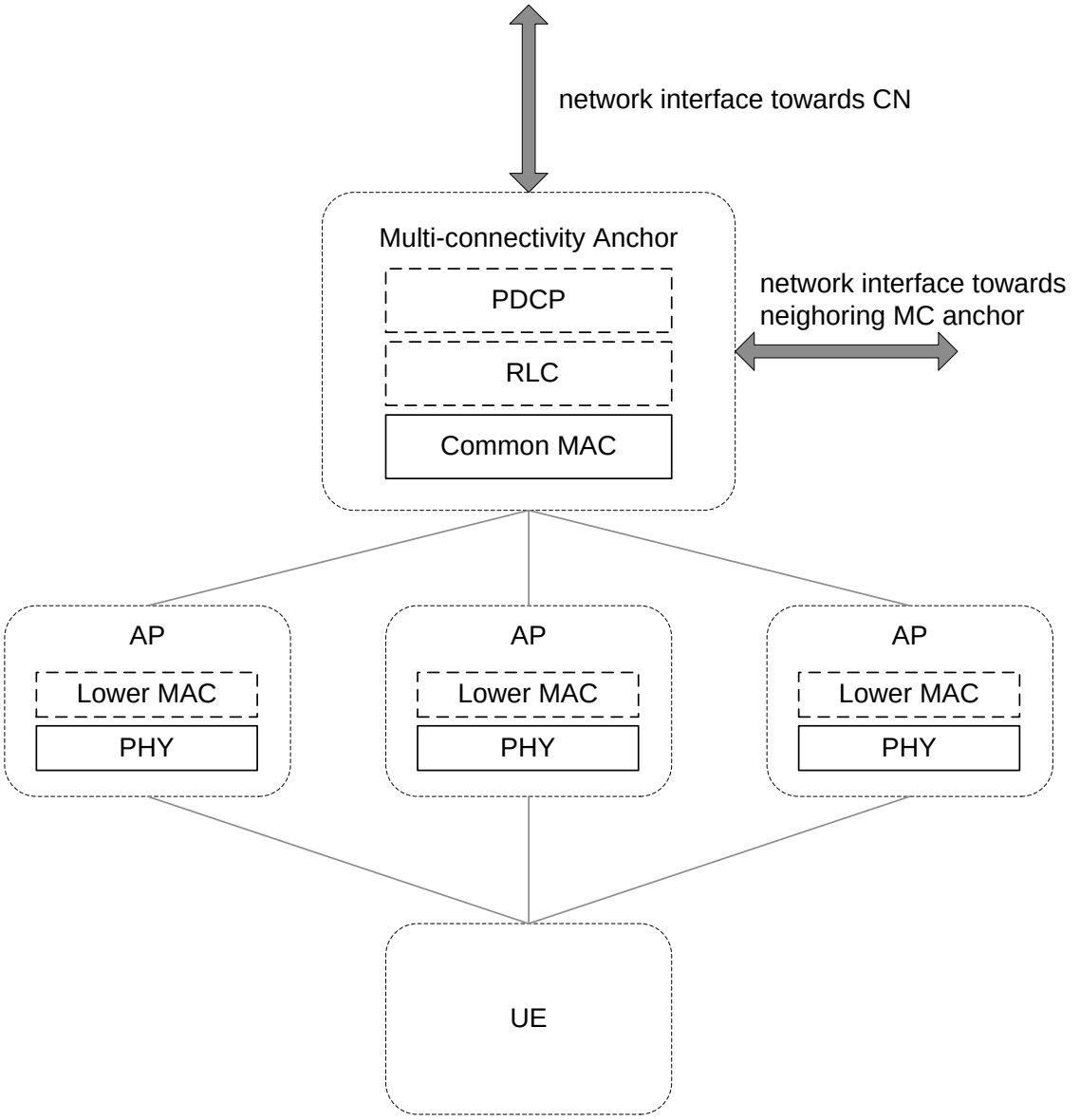


Figure 7-32: Summary and illustration of the proposal

It has been so far focusing on the new UL broadcast-based multi-receiver diversity MC scheme with the proposed CCMCC. The same or similar analogy may be used for introducing a single-frequency-network (SFN) based multi-transmitter diversity MC scheme for the DL direction. Figure 7-32 provides a summary as well as an overall illustration of the proposal.

Depending on the number and load status of individual APs in the current CCMCC as well as the quality of radio links between individual APs and the UE, MCN may select and configure a subset of the APs in the current CCMCC as mandated to monitor and receive targeted proximity broadcast based UL transmissions of the UE. This is referred to as the first subset of the said group stated above. The radio link quality between individual APs and UE may be either explicitly indicated by link quality measurement report from UE or implicit indication based on e.g. received and forwarded UL data by individual APs in both first and second subset group. For instances, if MCN detects that the UL data forwarded from one APs has occurred continuous error (e.g. based on CRC if data forwarding is on TB level or missed RLC SNs if data forwarding is on MAC SDU/RLC PDU level) while the other AP provides better radio link quality in term of corrected received UL data, MCN may reconfigure the both accordingly.

7.9. Multiple connectivity at the different layers,

7.9.1. State of the art

5G will have to cope with a diversity of access technologies and cell sizes (micro cells, Wi-Fi, mm Wave, among others). The user equipment will support different radio access technologies, and this flexibility will allow for better performance and reliability. Support for multi-connectivity should be an integral part of the 5G architecture. In this section, the different architectural requirements necessary for efficient support for multi-connectivity will be explored.

7.9.2. Current status of LTE

Even though LTE architecture was originally designed for voice and broadband services, it is quickly adapting to cope up with the future 5G requirements. LTE is also working on enhancing the multi-connectivity between different access technologies and heterogeneous networks. The investigation of multi-connectivity and the technology enhanced to achieve multi-connectivity in LTE is presented in the below section.

Dual Connectivity: The approach of multi-connectivity is standardised in 3GPP Release 12 with the launch of dual connectivity feature. Dual connectivity enables UE to connect with two different networks with different carrier frequency. One of the network nodes is the Master eNodeB and other is the secondary. The resources from two network nodes are assigned by two different RLC and hence, UE uses two distinct uplink carriers. The data layer of UE is connected to both network nodes; however, control signalling is carried out by backhaul X2 interface with single MME connection [RS-12]. The dual connectivity operates in two modes: Synchronous and Asynchronous mode. These modes are classified based on the delay spread that UE can survive.

LTE radio-WLAN Connectivity: The connectivity between LTE radio network and wireless nodes has been standardised from long time. The decision to select the best wireless connection was taken by core network with the introduction of Access Network Discovery and Selection function (ANDSF). ANSDF selects the wireless connection based on predefined rules that target offloading traffic from macro-cell with best QoS to the user. However, due to slow processing and non-existing support for ANDSF functionality in all UE, LTE introduces RAN Assistance Information in Release 12. The highest priority of selection is given to user to select the desired network. Secondly, ANDSF rules and RAN rules guides the network selection. If the ANDSF functionality is not supported by UE, the network selection is based on RAN rules. The RAN parameters (Received Signal Strength, WLAN Channel Utilisation, Available Backhaul Bandwidth etc.) information is measured by base-station and UE [RS-12]. These RAN parameters are compared with thresholds defined in RAN rules. The wireless network that satisfies these RAN rules is selected.

Heterogeneous Network Connectivity: LTE supports use of small cells to offload the traffic from base-station. The recent LTE release focus on optimisation of handover procedures along with interference mitigation between macro and small cells. The handover takes place considering the mobility of user. Users estimate and categorise its mobility as low, medium and high. Depending on the mobility information, base-stations estimate the time that the user will be under coverage of a small cell. The base-station then takes decision based on the mobility information, if the handover is feasible. Users also store history of 16 last visited cells [RS-12]. This historical information is provided to the base-station if requested and used to estimate the user's mobility. The other modification in the latest release was increasing the target cell specific Time to Trigger (TTT), that is, the time to send small cell measurement reports to the base-station [RS-12]. The modification prohibits handover to pico-cells if the user is moving with very high velocity.

License Assisted Access (LAA): As already discussed, LTE supports WLAN interconnectivity. The 3GPP Release 13 plans to increase the data rate and coverage by coordinating transmission on both licensed and unlicensed frequency bands. It also considers sharing of unlicensed frequency bands with other operators. LAA introduces 'Listen before talk' feature that enables it to

sense if selected carrier frequency is already in use for transmission [NOK-WPa]. The carrier aggregation technology introduced in Release 10 will be used in LAA. Release 12 has already standardised aggregation of TDD and FDD spectrum.

7.9.3. Current research on 5G

Insights on the architectural requirements for multi-connectivity can be retrieved by looking at some relevant proposed 5G architectures, current projects on 5G, and proposals for integration between multi-RATs.

In SoftNet [WH+15], radio access points in a unified RAN are connected with access servers at the edge of an SDN based core network. These access servers work as distributed mobility anchors, gateways and multi-RAT coordinators. Mobile network SDN controllers in [BOS+14] offer a southbound interface to the data layer entities in a unified RAN. Radio access points from different RATs are connected to a core transport backbone composed of programmable L2 switches and L3 routers. The access network is virtualised and hence allowing sharing of physical resources. Access technologies become programmable and are manage by the controllers to meet specific needs.

The work in [HN+14] investigates the benefits of multi-connectivity through a case study of integrating Wi-Fi with 3GPP. In the area of network architecture, the RAN will be responsible for distributing radio link information, either with UE assistance or through a defined interface between WLAN and 3GPP. The user will use 3GPP for transferring sessions from non-integrated cells, and then local switching for sessions to and from Wi-Fi in integrated cells. In the same area of integrating air interfaces, the work in [DMR+15] proposes a common integration layer, residing on top of the MAC layer of LTE and any new 5G air interface. This common PDCP/RRC layer for the control and data layers is considered a reasonable and future-proof choice.

In the METIS project [MET-D6.4], there are two high level building blocks directly concern with multi-connectivity. One is Radio Node Management (RNM), dealing with radio functionalities that affect more than one node. The other one is Air Interface (AI), handling air interface functionalities of radio nodes and devices. The former perform three important building block for multi-connectivity: RAT selection, radio resource management and interference management. AI contains building blocks directly enabling specific air interfaces.

7.9.4. Inter-RAT Integration Architecture

With the large acceptance of LTE and its heavy deployment worldwide, it is necessary to design 5G architecture that closely interworks with LTE. The integration of LTE with previous standards incorporate slow mechanisms, hence cannot be used for 5G-LTE integration. The transition from LTE to 5G is critical and will take time. Therefore, it is necessary to provide close integration of 5G with already existing standards, considering ultra-low latency 5G requirements, and simultaneously providing multi-connectivity to the devices.

Dual connectivity introduced in LTE Release 12 provides connection to two base stations, with different carrier frequencies but belonging to same radio access technology. With the launch of new 5G architecture, an interface between two base stations belonging to different access technology (e.g. LTE-5G, 5G-3G) is necessary. The current interworking architecture of LTE with previous standards (3G/2G) operates with very high latency, and does not incorporate ultra-low latency service requirement

The figure shown below is the general architecture for integration of different RATs. To provide RAT multi-connectivity to user equipment, UE must be simultaneously connected to base stations of different RAT's. The architecture is based on C-RAN, where all the radio network controller functions for 2G, 3G, LTE, Wi-Fi and 5G are integrated in to a edge cloud controller. The Edge controller is implemented with L1, L2 and L3 functionalities. The architecture has heterogeneous front haul network as it is connected to different Radio Resource Heads (RRH). The edge controller is connected to centralised core network with common interface for all RATs [WWRF].

The edge controller is responsible for inter RAT radio resource management, service mapping, inter-RAT interference mitigation and RAT selection. Based on the QoS/QoE requirements, and present network traffic, edge controller forwards the service request to corresponding core network. A service to UE can be provided by single RAT or multiple RATs, depending on the network load.

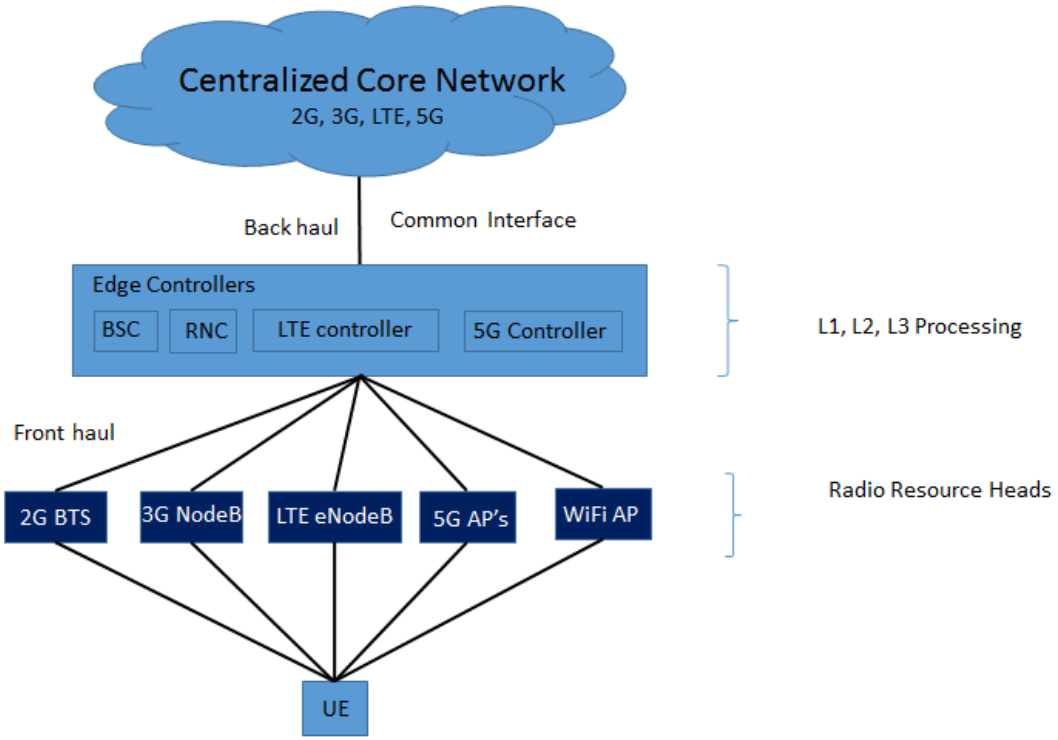


Figure 7-33: RAT Integration Architecture

In order to reduce the inter-RAT-working latency and signalling overhead of the system, a tight integration of 5G protocol layers with other RAT's layers is necessary. The recent research, propose use of single protocol layer for different RATs.

PDCP and RRC Layer: To reduce the signalling overhead, PDCP and RRC layer of 5G is shared for all the previous RATs. Therefore, the signalling messages can be sent through 5G RRH, even if the service is provided through LTE core network. The figure shown below is control layer protocol stack for RAT's integration. The control layer integration for Multi-RATs can be considered to be operating in two modes; to achieve reliability, and to reduce signalling overhead.

The first case, if the UE receives signalling messages through every RATs control layer. The signalling messages can be same or different over RATs control layer. This increases reliability of RRC messages, simultaneously allowing easy handovers. However, it increases the signalling overhead if same messages are sent multiple times. To reduce the signalling overhead new mode of control layer switching is suggested in [SMR+15]. The RRC messages are provided through single RAT control layer. The control layer is switched to anther RAT if required. The integration at PDCP and RRC layer is simpler and feasible, as it is asynchronous w.r.t TTI.

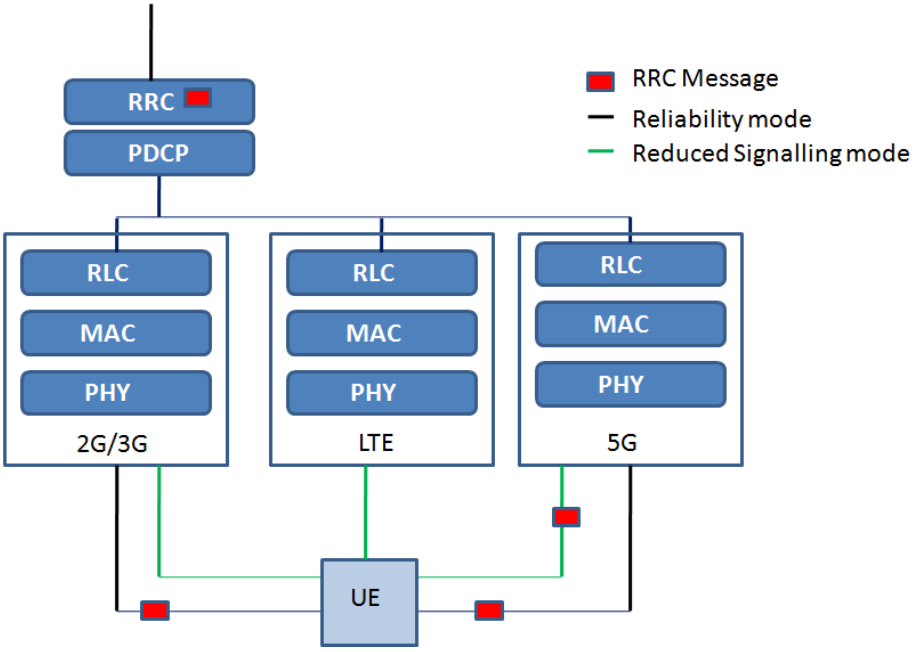


Figure 7-34: C-layer RAT Integration

Data layer: The throughput can be significantly increased if the single data flow is mapped over multiple RATs. The Flow Aggregation function is required to transfer the data packets for single flow over multiple RATs. The other function used is 'Flow Routing'. It allows single data flow to be mapped over single RAT. Multiple data flow per UE can be mapped over multiple RATs.

Also, common Multi-RAT MAC layer can give significant coordination and pooling gain, but requires very high synchronisation. [SMR+15].

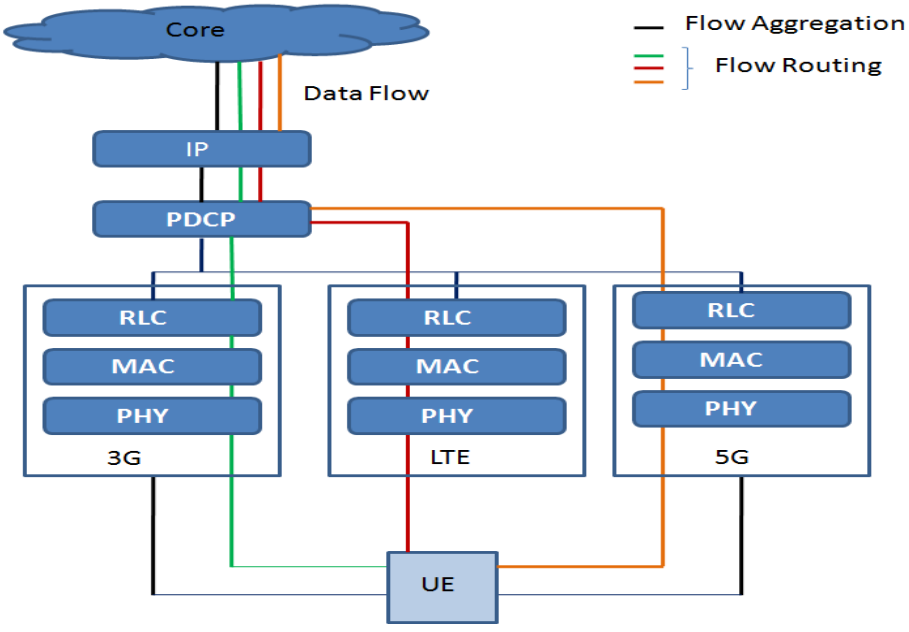


Figure 7-35: U-layer RAT Integration

7.10. Centralised Radio Resource Management

7.10.1. Introduction

The traffic properties of mobile networks have dramatically changed due to the proliferation of smart devices and traffic-hungry applications [CIS13]. In addition to rapid mobile data demand growth and the dramatic variation of traffic either geographically or temporally [KC15], the traffic symmetry in Uplink (UL) and Downlink (DL) has encountered extreme changes. As a solution, base stations densification was proposed to improve spectral and energy efficiency through enhanced control over coverage and interference [WHB15]. The cell-centric RAN is evolving to a new user-centric multi-tier architecture by implementing the small cells. C-RAN is the other key in the RAN evolution toward 5G. Coordinated Remote Radio Heads (RRHs), as the result of centralised based band processing offered by the C-RAN architecture, enables improvement of resource utilisation in addition to advanced features like enhanced Inter-Cell Interference Cancellation, Coordinated Multi-point transmission, and carrier aggregation [And13].

Hence, novel Radio Resource Management (RRM) approaches for the next generation of mobile network have to address various key topics such as dramatic variation of traffic pattern. Recently, traffic over mobile networks has shifted from symmetric voice-call dominant traffic (i.e., UL and DL resource consumptions are comparable [CFY04]) to burst-like traffic with severe fluctuation and resource exhaustion in one direction depending on the application type. These variations are expected to be more intense due to future implementation of 5G heterogeneous networks, where small cells are deployed in macro cell coverage.

The development of dynamic adaptation of TDD patterns is required in order to handle the rapid change of traffic patterns in the network. These approaches enable more flexible utilisation of the spectrum by assigning the resources of each frame to UL/DL dynamically with respect to traffic conditions in each cell. However, the cell-specific dynamic TDD pattern selection can lead to extreme UL/DL cross-link interference between neighbouring cells. Hence, approaches for interference mitigation have to be applied. The expected innovation in the framework of 5G NORMA for radio resource management can be summarised as selection of TDD patterns, mitigation of DL-UL interference, and overall improving radio resources utilisation.

7.10.2. State of the Art

According to the 3GPP specification TS 36.211, there are seven uplink-downlink configuration patterns for LTE-TDD, which offer different UL/DL ratios from approximately 60:40 to 10:90 within a system-frame consisting of 10 successive TTIs [PLS15]. These configurations are presented in Table 7-2, where “D” denotes downlink sub-frame, “U” uplink sub-frame, and “S” denotes a special sub-frame with three fields of DwPTS (Downlink Pilot Time Slot), GP (Guard Period), and UpTPS (Uplink Pilot Time Slot) [36.211]. Sub-frames number 0 and 5 are always reserved to be used in downlink direction. The UpPTS and the sub-frame immediately following the special sub-frames are always reserved for uplink.

Table 7-2: LTE TDD configuration (extracted from [36.211])

Configuration	Switch-point periodicity[ms]	Sub-frame Number									
		0	1	2	3	4	5	6	7	8	9
0	5	D	S	U	U	U	D	S	U	U	U
1	5	D	S	U	U	D	D	S	U	U	D
2	5	D	S	U	D	D	D	S	U	D	D
3	10	D	S	U	U	U	D	D	D	D	D
4	10	D	S	U	U	D	D	D	D	D	D
5	10	D	S	U	D	D	D	D	D	D	D
6	5	D	S	U	U	U	D	S	U	U	D

Furthermore, the support for faster reconfiguration of TDD sub-frames is introduced in LTE Release 12 [36.300]. The approach, which is known as "enhanced Interference Mitigation and Traffic Adaptation" (eIMTA), allows for dynamic adaptation of TDD patterns in response to varying capacity requirements in uplink and downlink. Any new scheme will be restricted to switching between these TDD patterns. For UEs supporting future eIMTA implementations, flexible sub-frames as illustrated by "F" in Table 7-3 are introduced. These can be configured dynamically either for uplink or for downlink. Legacy UEs will be configured with an uplink-heavy TDD configuration (in particular the "UL HARQ reference configuration", e.g. 0, see below). In other words, they are limited to uplink transmission in the flexible sub-frame. The eNodeB will not schedule a legacy UE with an uplink grant in case it wants to use the sub-frame for an eIMTA UE in the downlink. For the eIMTA UEs, an eNodeB can at best adapt its TDD pattern every 10ms; adaptations within a system frame are not possible.

Table 7-3: Effective eIMTA frame structure (extracted from [PLS15])

Sub-frame Number									
0	1	2	3	4	5	6	7	8	9
D	S	U	F	F	D	S/D	F	F	F

The most frequent method used is based on the uplink ratio of buffered traffic [SEK+12], where at first the percentage of uplink buffered traffic for each base station is determiend and then the pattern with the closest UL ratio to the calculated one is selected. In [LGL+13], the aforementioned approach is extended by adding the historical information as follows:

$$r_n = \beta \frac{R_{UL-req}}{R_{UL-req} + R_{DL-req}} + (1-\beta) \frac{R_{UL-his}}{R_{UL-his} + R_{DL-his}} \quad (7.10\text{-}1)$$

where:

- R_{UL-req}: overall data in the uplink buffer,
- R_{DL-req}: overall data in the downlink buffer,
- R_{UL-his}: total amount of the uplink data which has been transmitted during the predefined interval,
- R_{DL-his}: total amount of the downlink data which has been transmitted during the predefined interval,
- β: weight for balancing real-time data and history,

After determining the percentage of uplink ratio for each base station, the TDD pattern with the closest UL ratio to the calculated one is selected. A QoS aware dynamic uplink-downlink reconfiguration has been proposed, where the pattern selection for the cells are done based on both packet data rates and packet delay. In this paper, the network traffic is divided into two key category of Real Time (RT) and Non-Real Time (NRT).

$$E(i) = \exp\left(\frac{\alpha_i D_i[n] - \overline{\alpha D[n]}}{1 + \sqrt{\alpha D[n]}}\right) \quad (7.10\text{-}2)$$

where:

- α_i: delay constrain factor of user *i*,
- $D_i[n]$: delay of user *i* at slot *n*
- $\overline{\alpha D[n]}$: the average value in order to give more value to the RT traffic for TDD pattern selection.

7.10.3. Cell Clustering Concept

Cell clustering is a promising approach for configuring TDD-patterns in the network. The key idea is to group all the cells, which are close to each other and might cause interference in one cluster. For all the cells in the cluster the same TDD-pattern is configured. The clustering of cells can be done using the basic clustering approach, i.e., based on the proximity of the base stations

in a very rigid way, or it could be more dynamic and dependant on the current information of network. Therefore, there is still a lot of space for improvement in terms of optimizing the clustering and choosing the optimal TDD pattern.

Despite the advantages offered by the clustering approach, there are drawbacks too. One of these drawbacks is the significant reduction of network flexibility. Different cells in the same clusters can have very different DL/UL traffic ratios and using one TDD pattern would result in very inflexible solution. In [LWC15], authors proposed an algorithm called "Soft Reconfiguration" to increase the flexibility of the clustering by allowing each cell to have different TDD patterns even in the same cluster while controlling the induced interference. The algorithm runs in each cluster independently and in each iteration only one of the cells is allowed to change its TDD pattern to only the two closest ones in terms of DL/UL ratio. In the case of severe interference on the cell, reconfiguration process is triggered in order to decrease the interference. Using this approach, flexibility of clustering is improved while keeping interference suppressed.

In the [LCW15, LGL+13, DPC14], the clustering of the small cells is done based on the coupling loss between two base stations. Since the uplink is always more interfered, if the receiving power of the neighbouring base station at the specific base station is higher than the predefined value then they are merged into the cluster. Using this approach on the whole network, it is possible to split all base stations in pre-defined clusters. However, this approach is mostly static and it does not include the variations of channel conditions in the small cells.

Another interesting approach was used in [LGL+13] where the authors performed self-organizing algorithms using reinforcement learning and game theory, where small cells jointly estimated their time average performance and optimised their configurations. The objective was to minimise the inter-cell interference and maximizing the spectral efficiency of the network.

7.10.4. Centralised RRM for the Virtual Cells

In [SSP+16], the concept of virtual cell as a solution to pseudo-congestion (i.e., the phenomena where there are enough resources but in the opposite direction) has been proposed. The proposed approach enables dynamic frame alteration of each eNodeB in addition to allowing the UEs to use the available sub-frames of multiple eNodeBs.

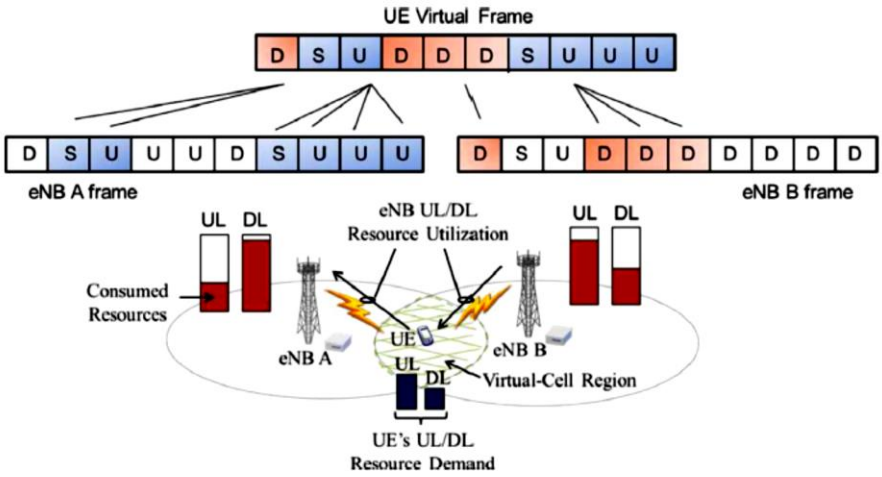


Figure 7-36: An example of virtual cell concept (extracted from [SSP+16])

The aforementioned concept of virtual cell is extended by detecting the edge-cell UEs and allowing them to use the neighbour cell resources in order to avoid inter-cell interference. Consequently, the total network throughput is expected to increase. Considering only the edge-UEs in resource allocation of virtual cells leads to faster overall scheduling. The key parts of the proposed

algorithm are a) detecting the congestion cells, b) detecting the cell-edge UEs, and c) alteration of TDD-frame patterns.

7.10.5. Detecting the Congested Cells

The first step in the proposed algorithm is to determine the congested cells. These cells can benefit from load balancing using the dynamic TDD pattern approach. There is a high correlation between PRB utilisation ratio and the PDCP packet delay, hence, the cell will be considered as congested if the following equation is satisfied:

$$\frac{\sum u_{BS}[n]}{N_s} > \mu_c \tag{7.10-3}$$

where:

- N_s: number of sub-frames in the TDD-frame,
- $u_{BS}[n]$: utilisation ratio in the n-th sub-frame,
- μ_c: the congestion threshold.

Furthermore, the cell is considered to be not congested if a similar expression holds true:

$$\frac{\sum u_{BS}[n]}{N_s} > \mu_{uc} \tag{7.10-4}$$

where:

- μ_{uc}: threshold for not congested cell.

7.10.6. Detecting the cell-edge UEs

Furthermore, after determining the congested cell, we also need to determine the cell edge-UEs in the cell which could offload DL or UL traffic in order to balance the TDD patterns. Hence, the user will be considered as the cell edge user if the following expression holds true:

$$|P_{MeNB}(ue) - P_{SeNB}(ue)| > \gamma_{DRP} \tag{7.10-5}$$

where:

- P_{MeNB}: received power from the strongest cell (Master eNodeB),
- P_{SeNB}: received power from the closest neighbouring cell (Secondry eNodeB),
- γ_{DRP}: the maximum difference of received power.

After determining the cell edge-UE, we also need to make sure if the neighbouring cell has enough resources to accommodate the user, as we don't want to congest the neighbouring cell. The following equation needs to be satisfied before assigning DL or UL connection of user to the neighbouring cell:

$$\alpha(SeNB) > r(ue) \tag{7.10-6}$$

where:

- $\alpha(SeNB)$: available resources in cell SeNodeB,
- $r(ue)$: requested resources of ue

7.10.7. Dynamic Alteration of TDD-Patterns

Finally, dynamic alteration of TDD-patterns is the final part of proposed algorithm. For each cell, the selected pattern has to meet the resource demands in UL/DL while the inter-cell interference is kept as minimum as possible. Regarding the inter-cell interference, the comparison of TDD-patterns is done based on Hamming distance, i.e., the difference between two TDD patterns is expressed as the total number of sub-frames where the TDD patterns are working in the opposite transmission slots. Hence, for each pattern:

$$TDD_x = \{SF_1, SF_2, \dots, SF_{N_s}\} \tag{7.10-7}$$

where SF_i: the i-th sub-frame and the value is 0 if is assigned to UL and 1 for DL.
The difference between the two TDD patterns of MeNodeB and SeNodeB is:

$$d(tdd_{MeNB}, tdd_{SeNB}) = \sum_{i}^{N_s} |tdd_{MeNB}(i), tdd_{SeNB}(i)| \quad (7.10\text{-}8)$$

where $d(tdd_{MeNB}, tdd_{SeNB})$ is the Hamming distance of the patterns assigned to MeNodeB and SeNodeB.

The pattern selection for each cells is done in a two-round procedure. First, it is assumed that all the UEs are connected to their strongest cell. Then, the required ratio of UL sub-frames to all sub-frames in each cell is calculated by:

$$r_n = \frac{R_{UL-req}}{R_{UL-req} + R_{DL-req}} \quad (7.10\text{-}9)$$

where:

- R_{UL-req}: overall data in the uplink buffer,
- R_{DL-req}: overall data in the downlink buffer,
- R_{UL-his}: total amount of the uplink data which has been transmitted during the predefined interval,
- R_{DL-his}: total amount of the downlink data which has been transmitted during the

For each cell, the proper TDD pattern is selected from the predefined patterns. Regarding the edge-cell UEs, they are assigned to the neighbour cell to reduce the load on the congested cell considering the congested direction in the MeNodeB. After assigning the edge cell UEs to relative cells, the traffic demands as is presented in (7.10-9) is recalculated and the relative patterns of the cells are updated.
In the following, the pseudo-code for proposed algorithm is presented:

1: Determine set of congested cells $C = \{c | \sum u(c) > \mu_c\}$
2: Determine set of un-congested cells $UC = \{c | \sum u(c) > \mu_c\}$
3: Create a set of pre-selected TDD patterns,
4: foreach $c \in C$
5: set of $UE = \{ue | ue$ connected to $c\}$,
6: **foreach** $ue \in UE$
7: $P_{MeNB}(ue) =$ received power of the strongest cell
8: $P_{SeNB}(ue) = received\ power\ of\ the\ closest\ neighbouring\ cell$
9: **if** $|P_{MeNB}(ue) - P_{SeNB}(ue)| > \gamma_{DRP}$
10: **if** $SeNB \in UC$ //check if the cell is not congested,
11: $\alpha(SeNB) =$ available resources in cell B,
12: $r(ue)$ = requested resources of ue
13: **if** $\alpha(SeNB) > r(ue)$
14: N_{MeNB}, N_{SeNB}:number of UL sub-frame
15: **if** $N_{SeNB} > N_{MeNB}$
16: pre-assign DL of ue to cell SeNodeB
17: else
18: pre-assign UL of ue to cell SeNodeB
19: end
20: recalculate the TDD pattern cells for both cells;

To this end, we are planning to implement the developed algorithm in the NOMOR emulator that allows the required flexible (de-)composition and allocation of radio resource management (RRM) in the context of heterogeneous networks by means of simulations for different network configuration, e.g. different backhaul delays [SMD15].

This serves to build a decision function which selects optimally functional distribution of the RRM algorithm depending on backhaul latency and service requirements (e.g. data rate and latency), within the mobile network, i.e. either near to the access network (decentralised) or remotely in the Network Cloud (centralised).

As next steps, we will perform a theoretical valuation of proposed centralised radio resource management algorithms and we extend the algorithm to allow the required flexible (de)composition and allocation of RAN functionalities (RRM) in the context of heterogeneous networks. The selected central algorithm is the evaluated by means of system level simulation for different backhaul delays. We will further design a decision function to select optimally functional distribution of the RRM algorithm depending on backhaul latency and service requirements, within the mobile network, i.e. either near to the access network ('edge cloud') or remotely in the Cloud (the 'network cloud'); we will identify service requirements (e.g. on data rate and latency) and deployment characteristics (e.g. frontal/backhaul latency) that impact the optimal placement of functions into network entities (quantitative KPIs); we will identify RRM functional splits to improve network scalability and flexibility (qualitative KPIs); and we will evaluate the requirements and constraints with respect to network virtualisation and RAN slicing [CSS+16, SSP+14, CSX+15]

7.11. Geolocation Databases, Use of Geolocation Information and Associated Opportunities

There have been extremely positive moves in the US, UK and Europe in general, Japan, Singapore, Kenya, South Africa, Tanzania, India, and elsewhere, towards the use of regulatory-based or regulatory authorised geolocation databases to better manage coexistence of spectrum users, prime initial examples including coexistence in TV bands through access to TVWS, in mobile communications bands through concepts such as Licensed-Shared Access (LSA), and in the 3.5 GHz "Innovation Band" in the US (see, e.g., [FCC10], [ITWS15], [FCC15], [GSM15]). These moves have been driven forward by pressure on spectrum, and notably the need to achieve spectrum sharing mechanisms in order to realise demand for mobile communications systems of the future—particularly at the low frequencies as will be necessary to underpin the novel mm-Wave developments in 5G, for reliability and other reasons. This need for spectrum sharing, along with ambitious concrete targets for the amount of "additional" spectrum that is realised by such sharing, has been enshrined in policy in the US and Europe (see, e.g., [PCAST12], [ECC15]).

It is noted that some early-movers on the capabilities that will be realised through such databases, are also involved, to various extents, in the development of equipment [TWSD15]. Moreover, as long as these databases have an agreement with the equipment operators that they are serving, they may also "manage" that equipment using the capabilities that are created by concepts such as TVWS, among others. In the UK and European TVWS rules [ETSI14], for example, such geolocation databases must maintain information on the unique identities of devices, their technical capabilities including particularly spectrum mask class, their locations (including height above ground level) and other characteristics, and all of the signalling and other technical capabilities are present and proven in order to keep such detail. Moreover, these databases also access and use, for propagation calculations, detailed information on aspects such as building clutter and propagation variations in a given location. Such a wealth of information and capability at a given management point leads to great opportunity to use the information for other purposes aside from merely allowing license-exempt opportunistic secondary spectrum access, as has been done in the case of TVWS. Moreover, it is noted that these databases and exactly the same procedures can be twinned also with licensed access, as is already planned in the UK in TV White Spaces [MCWSD15].

The prime target of 5G NORMA is clearly not spectrum or spectrum sharing. Moreover, systems are already heavily under development or otherwise realised for connectivity among users in particular domains or for particular types of systems. However, in order for the wealth of heterogeneity to be maximally realised, also taking advantage of the vast potential for additional mobile communications resource access through TVWS, LSA, and other sharing realms, it is necessary

to bring such higher-level and third-party authorised databases into the equation. This, along with their strong potential to facilitate other forms of rendezvous including connectivity awareness among the vast range of heterogeneous access and network types feeding 5G capabilities, brings geolocation databases into scope as an option for such connectivity awareness as a management point. Moreover, for optimal operation and to minimise latency, such databases should ideally be realised in a virtualised form, using cloud-based processing. This is already the case for US 3.5 GHz "Spectrum Access System" developments, for example. Moreover, the information in such databases can be built or greatly assisted using sensing mechanisms, as already is the case in the US 3G GHz SAS. Further, sensing to assist connectivity awareness has been realised in the IEEE 1900.6a standard, for example, and the combination of such sensing information with spectrum databases is further being dealt with in the ongoing work on the IEEE 1900.6b standard [IEEE1900.6].

Given such opportunities, using exactly the same information that is already present in geolocation databases under the UK and European TVWS framework, the databases might:

1. Manage the resource usage also among the opportunistic users that are accessing the spectrum.
2. Linked to this, be reasonably aware, through propagation calculations between the users given knowledge of their locations and knowledge of propagation characteristics in given locations and associated information, of which users could potentially connect in a heterogeneous access environment. These users might otherwise be unaware of each other due to their different domains of operation and, e.g., their automated or manual use of power saving mechanisms to save battery and energy.
   The databases might also allow information to be conveyed on location maps of connectivity options, allowing the 5G heterogeneous network to plan into the future.

Given such observations, it is suggested here that such an integrated system for coordination supporting 5G is based on these databases, due to: (i) the involvement of the regulator in the process where the full range of resource usage opportunities is concerned, hence the presence of a higher-layer databases with information and capability to act in such roles, (ii) the establishment and trialling of them in various contexts and likely further building on such concepts in regulatory and other circles, and (iii) the advancement of such approaches (e.g., LSA) to support mobile communications cases. Moreover, it is suggested that these databases should be of three forms: (i) a form of database in the presence of and run by the regulator, (ii) a form of database trusted and likely certified by the regulator but existing and being run outside of the scope of the regulator, and (iii) a final (optional) form of database that is untrusted, and existing and being run outside of the regulator.

Figure 7-37 gives a representation of how such a hierarchy of databases might look. Detailed discussion on the reasoning for this is given in [HD15].

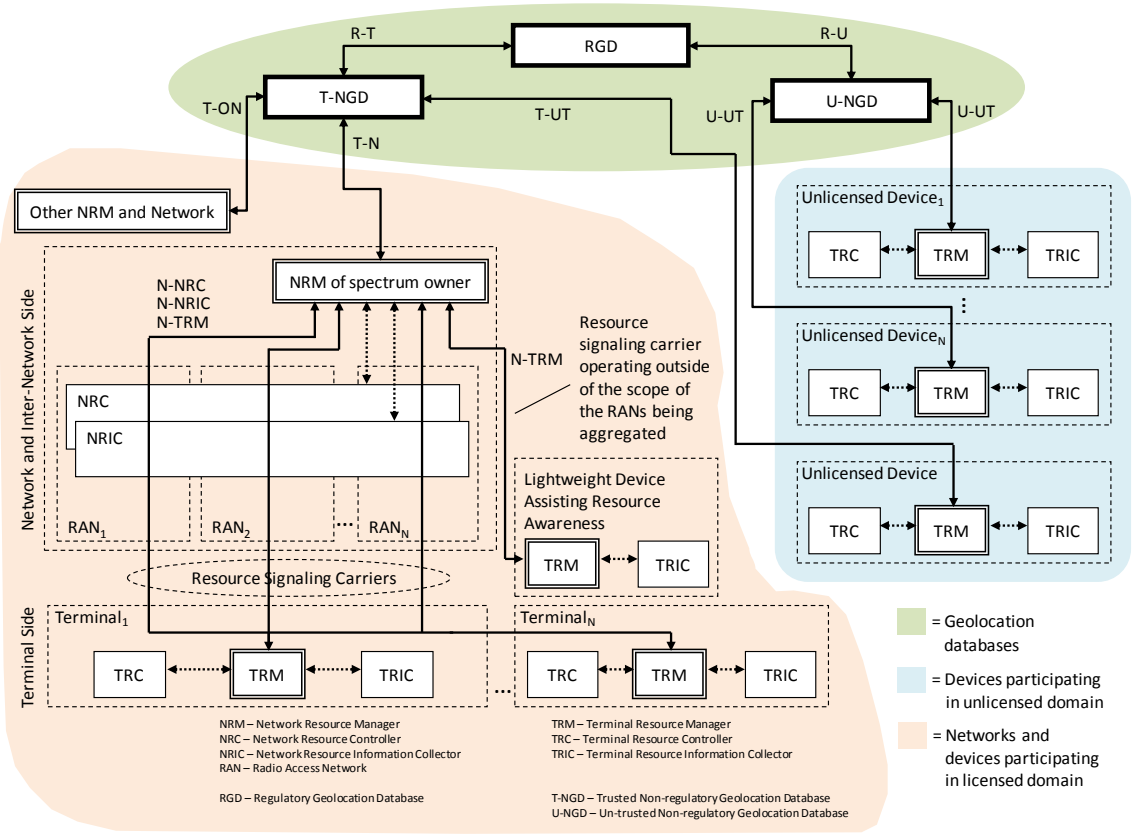


Figure 7-37: Example of the form that such a geolocation database-based architecture might take

Further, it is noted that in addition to the above applications, geolocation is needed due to the nature of the link requirements between source and destination, e.g., to minimise latency though optimising the propagation path based on geolocation information. Figure 7-38 illustrates an example of this through the realisation of a virtualised optimised propagation path for the Tactile Internet as a key 5G and beyond application, supported by signalling over the mobile network for control purposes [HWFGQ15]. In such cases, the geolocation databases themselves might often need to be virtualised and optimally located in order to minimise latency in signalling with the geolocation databases in order to make geolocation-based decisions.

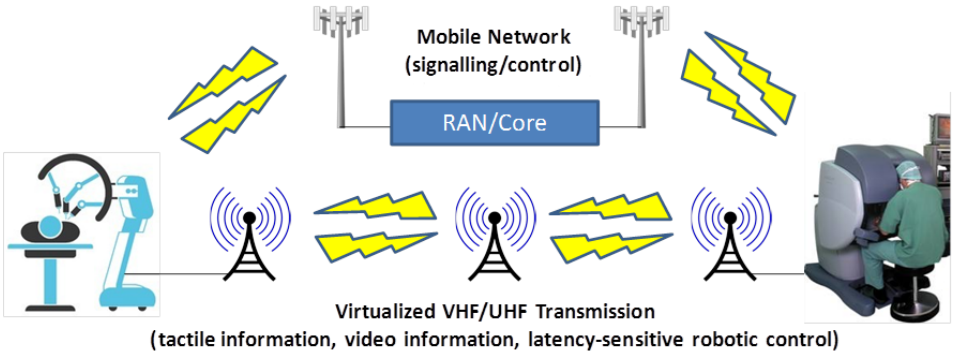


Figure 7-38: Illustration of Tactile/Haptic Internet/Communications and virtualised shortest-path links for latency minimisation

7.12. Multi-Tenancy in Multi-RAT Environments

5G NORMA follows the objective to extend the multi-tenancy concept also to Multi-RAT networks. In this case our goal is to share not only the resources of different BSs, but also to share the different technology. For example, we have one BS and two different radio access technology (Wi-Fi, cellular network like LTE). We could decide to provide to all users of a particular tenant only the Wi-Fi resources and use the cellular ones for the users of other tenants. In this scenario, we have more allocation possibilities so the problem is more complex and also the algorithm needs to be redesigned in order to fit the 5G requirements.

7.12.1. Network sharing in 3GPP

3GPP mainly focused on the definition of new sharing scenarios and requirements, and the corresponding network management architectural and functionality extensions toward on-demand capacity brokering. 3GPP provides in [32.130] two different RAN scenarios. First, RAN-only sharing (MOCN) where different tenants share only the RAN elements and secondly, Gateway Core Network (GWCN) where different tenants share not only the RAN elements but also part or all of the Core Network elements.

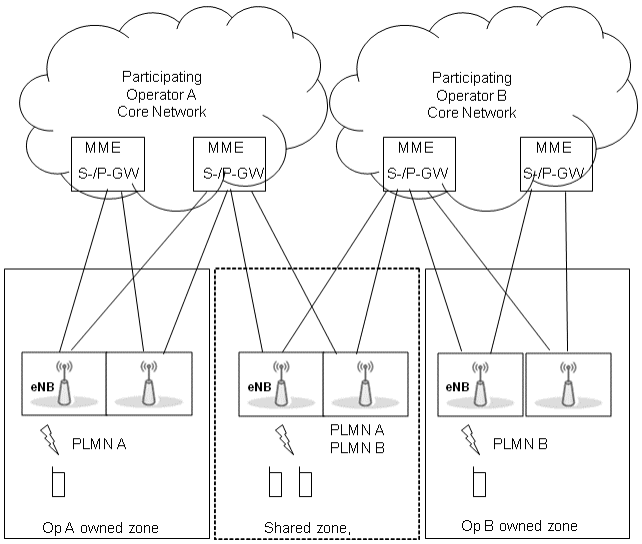


Figure 7-39: Multiple Operator Core Network (MOCN)

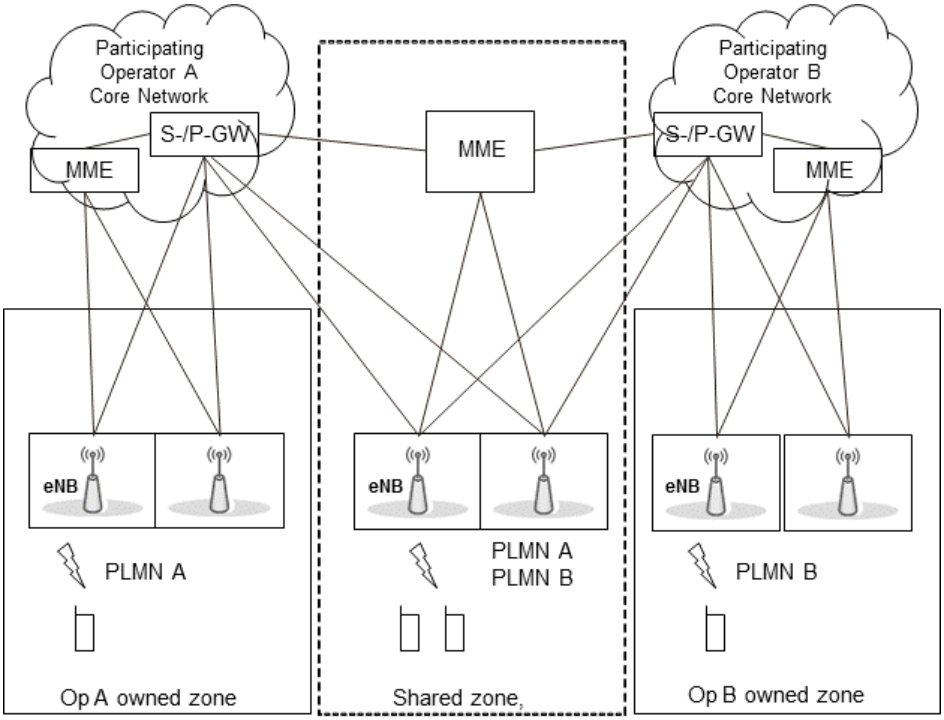


Figure 7-40: GateWay Core Network (GWCN)

3GPP also defines the management architecture, the requirements for OAM&P and the Actor Roles:

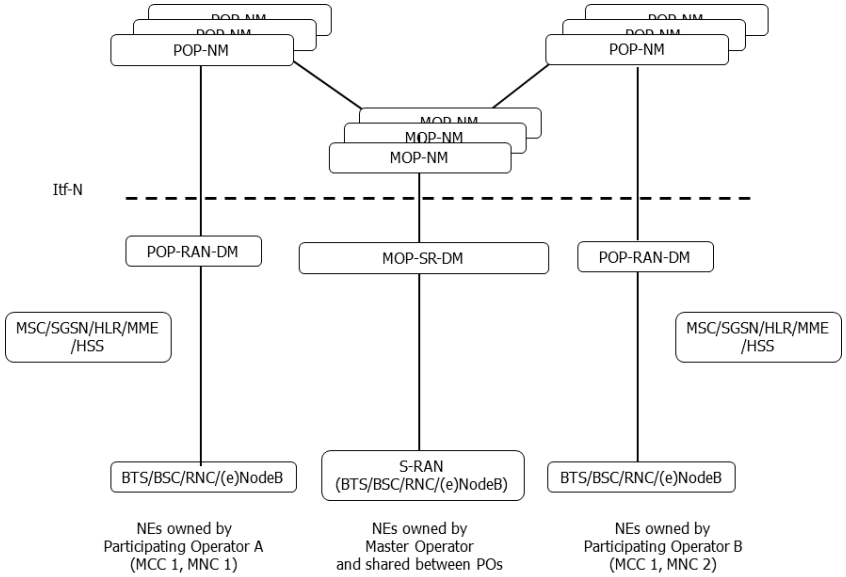


Figure 7-41: Management architecture for MOCN

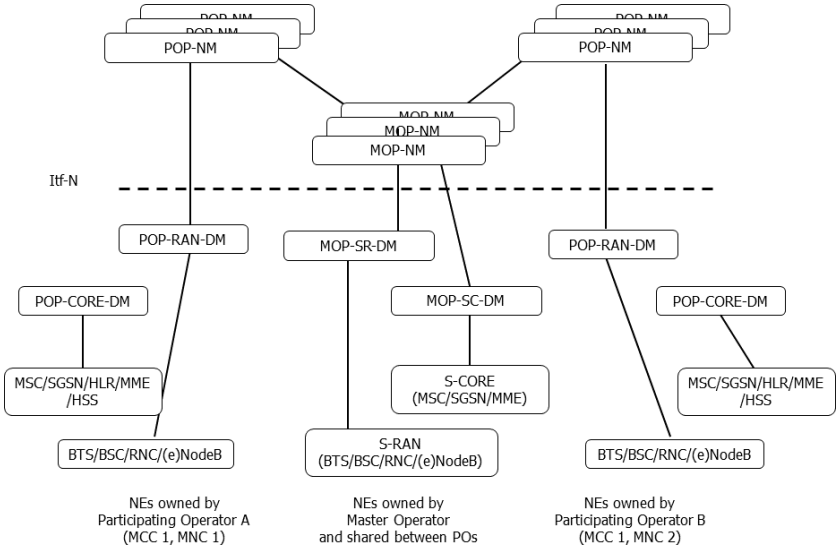


Figure 7-42: Management architecture for GWCN

In [22.101], 3GPP specifies the sharing requirements for E-UTRAN and similarly for GERAN and UTRAN. When E-UTRAN resources are shared they can be allocated unequally to the Participating Operators, depending on the planned or current needs of these operators and based on service agreements with the Hosting E-UTRAN Operator.

The following requirements apply: the Hosting E-UTRAN Operator shall be able to specify the allocation of E-UTRAN resources to each of the Participating Operators by the following:

a) Static allocation, i.e. guaranteeing a minimum allocation and limiting to a maximum allocation,
b) Static allocation for a specified period of time and/or specific cells/sectors,
c) First UE come first UE served allocation.

Resources include both data layer and signalling layer. The Hosting E-UTRAN Operator needs to be able to manage the sharing of the signalling traffic independently from that of the user traffic because signalling traffic and user traffic are not always directly related.

The management and allocation of resources of signalling traffic over the Shared E-UTRAN shall be independent from the management and allocation of resources of the user traffic over the Shared E-UTRAN.

3GPP provide also a procedure in situations where a need for additional, unplanned, E-UTRAN capacity by a Participating Operator arises (e.g. in the case of big mass-events). In this case the Shared E-UTRAN can provide means to allocate available spare capacity to the Participating Operator. Based on service level agreement between the Hosting and Participating operator such allocation can be automated without human intervention.

The Participating Operator shall be able to query and request spare capacity of the Shared E-UTRAN, based on policies and without human intervention.

The Shared E-UTRAN shall be able to allocate spare capacity to Participating Operator, based on policies and without human intervention.

As we can see the specifications provide by 3GPP are different from our vision of multi-tenancy and they don’t cover multi-tenancy in Multi-RAT scenarios. This is the reason that move us to design a new functionality that enable the 5G vision.

7.12.2. New multiplexing ability

Substantial attention has been devoted to the architectural framework for multi-tenancy, but relatively little work has focused on criteria/algorithms and state-of-the art ones fail to meet the requirements for a practical solution or they have been proposed without proper justification.

In order to extend the Network Sharing 3GPP standardisation to meet the 5G requirements, we need to design:

- New sharing criterion (see Section 7.5)
- New multiplexing ability

The new multiplexing ability has to be designed based on the above criterion. For that reason, we focus on a more dynamic sharing concept, signalling-based and with no human intervention, which enables a more efficient sharing of the network resources according to SLA required and taking into account also the commercial agreement.

Given the amount of information involved (including the channel quality of each user) and its dynamic nature, the algorithm should be distributed. Also, since the algorithm may be triggered frequently (whenever a user joins, leaves or changes its location), it should be computationally efficient. When adapting to network changes, the algorithm should control the number of handoffs triggered, as those may represent high overhead.

We have to possible way to implement it:

1. One new multiplexer function with a scheduler per each network slices;
2. One scheduler for the network that allocates resources among different slices.

In the second case we don't need a new function because we could implement directly the algorithm in the scheduler. In the first case the scheduler has to cooperate in order to obtain an optimise allocation of the resources overall network.

The new ability has to work also in Multi-RAT scenarios. In this case we need to consider more possible allocation solution because we share not only the resources of each BSs but we can decide to share also the different technologies of each BSs among all users. The problem now is obviously more complex.

We can conclude that at the state-of-the-art all efforts concerned multi-tenancy are more focused on architectural and requirements aspects. So we need to design a completely new scheduling/multiplexing functions with new algorithm to allow the new 5G functions both for multi-tenancy in normal RAN and multi-RAT scenarios.

7.13. Load Balancing of Signalling Traffic

The control and user (i.e., data forwarding) layer separation (C/U split) is expected to provide a significant facilitation towards the co-existence of different radio technologies including high capacity small cells within the same network that are orchestrated by a common control infrastructure via macro cells. Such a re-thinking on the cellular architecture has been propelled by requirement to accommodate future needs; it is now estimated that by 2020, mobile user demand for data will generate an order of thousand times the traffic level on mobile networks compared to year 2010. In a c/d split network domain macro cells act as the network control layer that efficiently manage in real time resources of a large number of high capacity small cells (including future mmWave-based cells, aka 5G) that can serve mobile users "on demand". To handle the increasing data demand from mobile users it comes as no surprise the move towards the use of very dense, low-power, small cell wireless networks that will unlock the potential of extremely high spatial reuse (especially when utilizing spectrum at the mmWave bands). Small cell networks allow for increased capacity by spatial localised transmissions (i.e., bring small access points closer to the user) whilst leveraging a more efficient spectrum utilisation by allowing multiple concurrent connections to different access points for a mobile user [NGMN15]. The concept of C/U split has been envisioned in the setting of small so-called phantom cells that will serve mobile users (D-

layer) while being controlled by a macro base station operating at different (lower) frequency bands [RWS12], [5GWP14].

Due to the high density of small cells, macro base station and small cell association and coordinated management will become an increasingly difficult and complex network function in the near future. In that domain we aim to detail a set of optimisation problems aiming to increase the performance of C/U split architectures under the assumption of a potentially large number of small cell and associated user mobility. More specifically, we are interested in dynamic, on-line, policies for assigning small cells to macro base stations controllers by taking into account control layer load conditions aiming to avoid degradation on the performance due to network congestion episodes. Within that setting, we furthermore assume a logical decoupling between the control and data layer domain via software defined networking (SDN). In essence, SDN enabled wireless access and core networks provide a hardware agnostic programmable framework for easing the development of new network functionalities and isolating complexities through the separation of control and data layer [GPL+13]. SDN will allow the required flexibility in network monitoring, policy installation and network management and hence act as a catalyst in allowing novel network orchestration techniques to be adopted in the c/d-layer split wireless architectures as envisioned in this paper. In addition to that, advanced algorithm will be able to run via network cloudification [BCJ+12], fully programmable RAN [BMK+12] as well as sharing between operators or third parties of the physical network infrastructure [MKH+13]. Load balancing has been mainly considered in macro cell mobile networks via various versions of the "cell breathing" technique [BH09] where BSs adjust their coverage area by changing their transmission power depending on load conditions. Recently some efforts have been detailed in load balancing in HetNet environments [CBR11], [HYP+12] which are more relevant to the current work of 5G small cells but to the best of our knowledge there is very little work for load balancing in C/U split architectures.

Under the above described framework, we will be assuming a future high dense network scenario with a cell density of at least ten times higher compared to current levels as expected in 5G networks and in addition to the pico cells we also assume a number of macro-controllers that have overlapping coverage areas. The proposed set of optimisation problems relate to the cases where a pico cell can be served by a number of different macro controllers. We aim to minimise overall network control overhead by assigning pico cells to macro-controllers in a way that reduce required handover traffic using historical data on handover rates between different pico cells. This is a pragmatic assumption since such information is readily available from mobile network operators.

7.13.1. Load Balancing of UE Agents

In section 7.2, we have outlined the optimised on-demand RAN level decomposition of E2E connections approach that shall be enhanced with above described framework to obtain the balance of usage and optimum performance of the network. The concept of the UE Agent and given the rationale for its existence in terms of achievable gain as well as required signalling, in order to realise the envisioned services. The means for balancing the load of different UE Agents so that the overall system performance is increased, up to a certain maximum number of UE Agents to be accommodated in the network. To visualise the problem, the illustration in Figure 7-43 presents the case where a number of UE Agents [SPW+96] need to be installed/located/distributed on different nodes in the network, in order to serve users.

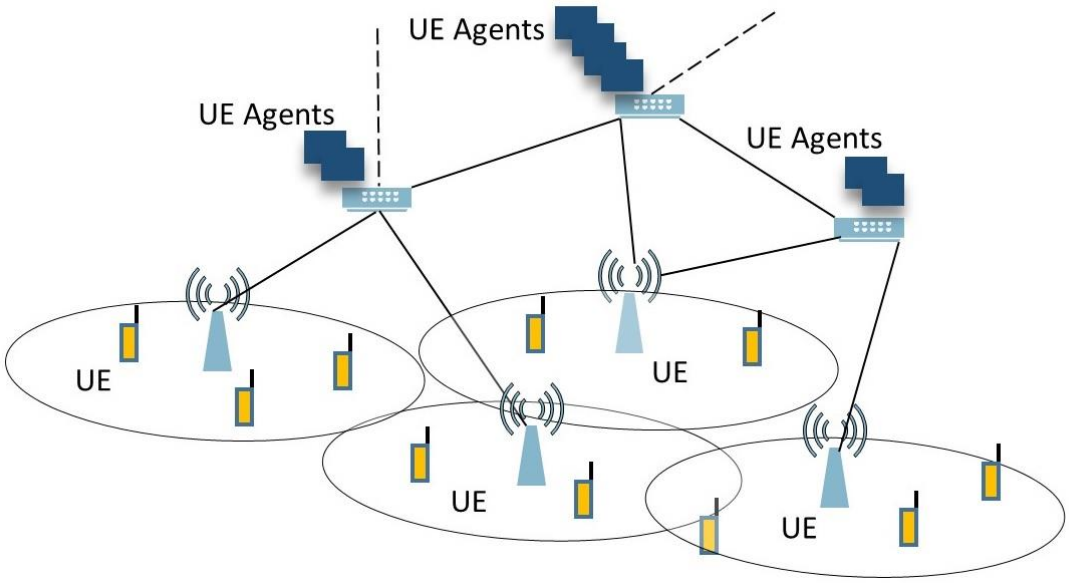


Figure 7-43: Load balancing in the network nodes that host UE Agents in order to allow increased level performance

In this scenario, it is important to balance the load among the different network nodes which can host the UE Agents. To this end, we can provide an association of UE Agents to network nodes such that we achieve load balancing across the various nodes in the networks by ensuring at the same time that node capacity is also satisfied (i.e. the number of UE Agents and network nodes D and M respectively and we define a binary decision variable:

$$x_{ik} = \begin{cases} 1, & UE\ Agent\ i\ hosted\ on\ node\ k, \\ 0, & otherwise. \end{cases} \tag{7.13-1}$$

Using the above definition, the UE Agent load balancing problem can be formulated as

$$[Prob\ 1] \min \sum_{k \in M} \left[\sum g_i x_{ik} \right]^2,$$

$$s.t. \begin{cases} \sum_{i \in D} g_i x_{ik} \leq C_k, \forall k \in M, \\ \sum_{k \in K} x_{ik} = 1, \forall i \in D, \\ x_{ik} \in \{0,1\}, \forall i \in D, \forall k \in M \end{cases} \tag{7.13-2}$$

where:

- g_i: the size of the content/file that UE Agent i is required to cache,
- C_k:the available storage capacity for each netowrk node k

The above equations explain a network snapshot that has not been optimised. Constraint of this formulation ensures that each UE Agent will be hosted in only one network node. Note, however, that the problem is not linear and therefore we cannot apply integer linear programming techniques directly. To resolve this, the problem can be reformulated in order to be made linear, as:

$$[Prob\ 2] \max t, \tag{7.13-3}$$

$$s.t. \begin{cases} \sum_{i \in D} g_i x_{ik}, \forall k \in M \geq t, \\ \sum_{k \in K} g_i x_{ik} \leq C_k, \forall i \in D, , \\ x_{ik} \in \{0,1\}, \forall i \in D, \forall k \in M \end{cases}$$

The above problem can be efficiently solved using integer linear programming techniques, though for large instances heruistic techiques will need to be implemented since the complexity of integer programs is not amenable to polynomial time solvability.

7.13.2. Load Balancing Numerical Investigations

This section illustrates a set of insights into achievable performance and improvement via simulations of uploading and downloading large contents with load balancing of UE Agents in the network. Table 7-4 indicates the simulation's parameters and values, the simulation results given in Figure 7-44 show that the maximum load of network nodes hosting UE Agents has been reduced significantly. The improvement that the proposed optimisation framework can achieve can be estimated as on average around 25%, among the different scenarios.

Table 7-4: Simulation Parameters

Parameters	Values
Number of network nodes (M)	3
Number of UE Agents (D)	5, 10, 15, 20
UE Agent requirements	0.6 – 0.9 Mbps
Network node available capacity	9 – 29 Mbps
Congestion Level	0.4 – 0.9

This load balancing framework introducing UE Agents close to the end user, such that overall upload/download performance can be improved. The load balancing scheme pertains to the associations of the UE Agents with the nodes in the network. The analytical and numerical/simulation results that load balancing scheme significantly reduces the peak load among the nodes in the network, and that the UE Agents scheme in general improves upload/download performance.

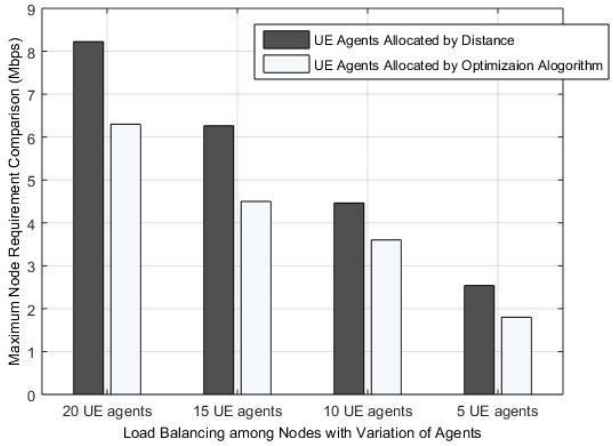


Figure 7-44: Load balancing of different network nodes based on the requirements of UE agents

Advantages

The key advantage of the proposed approach is that it allows for consistently providing optimal solutions to both overhead reduction problem and load congestion (balancing) in the control layer traffic. In that respect, it provides an upper bound indicator on the performance improvement that can be achieved in the network.

Disadvantages, further issues to be considered

However, there are still some disadvantages need to be raised up. First of all is the practical consideration of testing the algorithms. The experimental results are based on extended Monte Carlo simulations using MATLAB but care should be taken to translate those with respect to more realistic network cases in terms for example small cell deployments and actual users handovers that depend on their mobility pattern. Clearly, those can vary significantly since they depend very much on the location and user distribution.

The proposed algorithm per se requires **average handover rates** between adjacent cells to be monitored and being readily available to the algorithm and **cell load factors**. This type of information will need to be communicated from the small cells to the network controller (and being periodically updated). The evaluation was based on assumptions regarding the values of those metrics but to get a more realistic set of results, the algorithms will need to be tested on real world traces. For example, average user handover rates between adjacent cells could be acquired from network operators but it is in general difficult to have access to such data especially for small cells deployments of high density.

Furthermore, the proposed mathematical programming setting is in general not a scalable framework due to the inherent non-polynomial complexity of the problem at hand. Numerical investigations reveal that it is possible to run such a framework for low to medium network size instances but might lead to increased running times for large network instances. Therefore, heuristics and/or greedy solution methodologies need to be defined in order to allow scale free operation for any network instance. One issue that also needs to be considered is if there are any conflicts between the proposed algorithms and other network functions in network entities (such as for example admission control, even though this functionality is not being investigated within 5G NORMA). These issues will be more clearly shown and/or clarified via a signalling procedure that will detail how this framework can be considered as a building block of a generic SDMC controller in the network.

7.13.3. Load Balancing of Cross-Optimisation Challenges

The key cross-issues between mobility, multi-path routing and VNF chaining and routing. The aim is to illustrate some generic cases and propel the idea that a joint design is required in order to optimise the performance of a system architecture similar to 5G NORMA.

7.13.4. VNF Location and Chaining Problem

The VNF location and chaining problem requires to find the optimal number and placement of VNFs by taking into account the number of service requests and the ordering of the visiting VNFs in order to minimise the overall network operational costs and the utilisation of the network. Depending on the nature of VNFs some more detailed constraints will need to be taken into account such as for example the anti-affinity constraint which requires some VNF that provide a correlated access the physical underlying resources to be implemented in different physical nodes because isolation of the operation of the different VMs might be compromised. This problem is also called the service chaining problem which, as mentioned above, relates to the process of routing a network flow (service) over a number of NFs in a pre-defined order. The service chaining problem can be easily shown to fall into a special category of facility location problems which are in general NP-hard and therefore intractable for pseudo-real time solutions of large network instances

[CB10, MSG+15]. OpenNF proposes an implementation of the control layer for VNFs as well as the network data layer by extending SDN functionalities [GVP+14].

Despite the significant attention that this problem has received over the last few years, there has been very little work on its application in wireless mobile networks. This is especially true with respect to the issues of mobility, QoS and multi-path routing that could provide the means for service differentiation and increased levels of network utilisation.

7.13.5. VNF Routing & Chaining with Mobility Support

Most of previous proposed solutions on the issue of VNF chaining and routing do not take into account user mobility. When user mobility is taken into account, we can potentially have the case where the path between the node of the last-in-order VNF to be visited and the service access router is changed due to a handover to a different access router. Moving to a new access router means that the above last routing path segment will be changed and therefore the chaining and location of VNF might not entail optimal operation for the network. Hence mobility issues need to be taken into account and potentially a joint optimisation scheme should be implemented that takes into account the effect of service migration/handover to a different access router. The above joint design could be implemented in per-network flow basis or for aggregate network flows using statistical information by exploring historical data on aggregate number of handovers in the specific geographical location.

7.13.6. VNF Routing & Chaining with Multi-Path Support

The envisioned decoupling between control layer and data forwarding plane via SDN allows for incorporating novel and flexible routing schemes compared to the current approaches, which are mainly based on a single shortest path between two communicating network entities in the network. To this end, multi-path routing is a feature of high promise which has yet to be explored in emerging architectures and an above mentioned programmable forwarding plane and control layer will propel such solutions from concepts to real-world implementations.

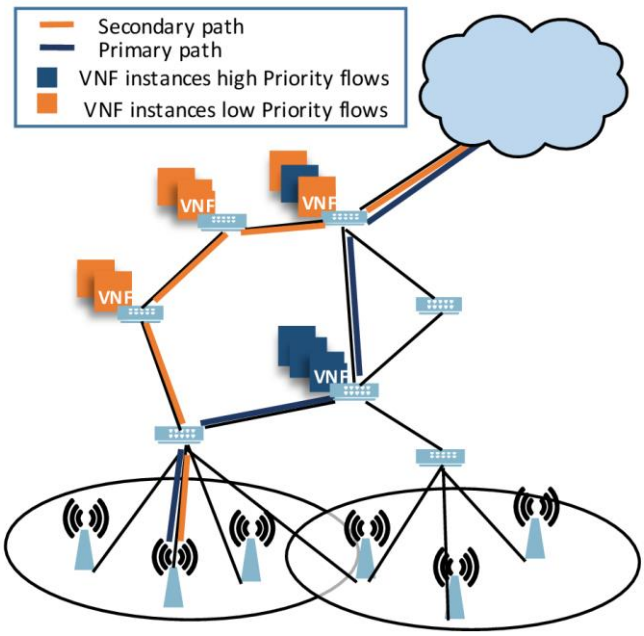


Figure 7-45: An illustration of VNF chaining and routing using multi-path routing for service differentiation and better utilisation of network resources

As shown in Figure 7-45, high priority service flows are allowed to use the shortest path (blue lines) and perform VNF chaining whereas low-priority service flows can utilise a secondary shortest path (orange line) to perform VNF chaining (see Figure 7-45).

Multi path routing can be utilised for inter-VNF routing to provide efficient utilisation of available resources and to provide policies for per-flow treatment based on different service flow priorities. An example of that scenario is depicted in Figure 7-45, which shows the case of two flows with different priorities and how multi-path routing can be jointly executed with VNF routing and chaining so that network resources are better utilised. In a more general framework VNF chaining and routing can be composed in order to fulfill Quality of Service (QoS) as well as Quality of Experience (QoE) constraints and/or requirements.

7.14. QoS innovation

7.14.1. State of the art

QoS can be defined as a set of characteristics related to the performance of the elements that provide the services that have an effect into final end users' perception.

3GPP has defined the LTE QoS using the Evolved Packet System (EPS) bearer model and is implemented between UE and PDN Gateway. A bearer can be seen basically as a virtual concept and is a set of network configuration to provide special treatment to set of traffic e.g. VoIP packets are prioritised by network compared to web browser traffic. In LTE, QoS is applied on the EPS bearer that is composed of multiple element bearers, Radio bearer, S1 bearer and S5/S8 bearer.

- Radio bearer – carries the packets of an EPS bearer between the UE and the eNodeB.
- S1 bearer – carries the packets between the eNodeB and the Serving Gateway (S-GW).
- S5/S8 bearer – transports the packets between the S-GW and P-GW. In principle S5 and S8 is the same interface, the difference being that S8 is used when roaming between different operators while S5 is network internal.

Each bearer (user data) path in LTE is assigned a set of QoS criteria and services with different QoS criteria need additional bearer paths.

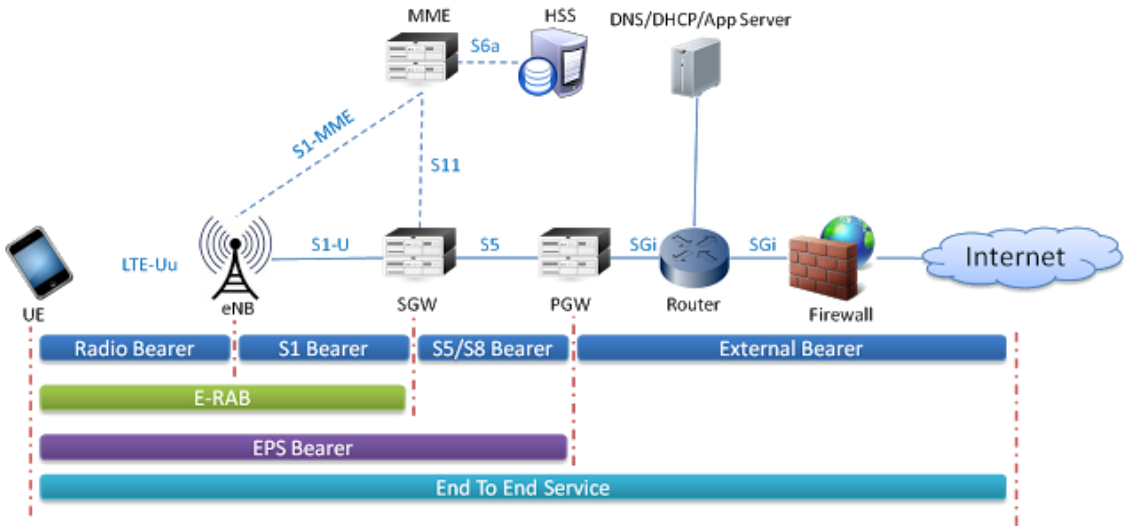


Figure 7-46: Radio Bearers [4GLTE13]

EPS Bearer

A bearer is a basic traffic separation element that enables differentiated treatment for traffic with different QoS requirements and provides a logical, edge-to-edge transmission path with defined QoS between the user equipment (UE) and packet data network gateway (PDN-GW).

Each bearer is associated with a set of QoS parameters that describe the properties of the transport channel. All flows mapped to a single bearer receive the same packet-forwarding treatment.

The QoS parameters associated to a bearer are:

- QoS Class Identifier (QCI)
- Allocation and Retention Priority (ARP)
- Guaranteed Bit Rate (GBR) (Real time services only)
- Maximum Bit Rate (MBR) (Real time services only)

There are two types of bearers, default and dedicated bearer.

Default bearer. It is assigned when a mobile device is attached to an LTE network and remains as long as UE is attached. Each default bearer is associated with an IP address, UE´s IP address. The default bearer provides only best-effort service, non-guaranteed bit rate (non-GBR), and an UE can have more than one default bearer. Each default bearer will have a different IP address. QCI 5 to 9 (Non- GBR) can be assigned to default bearer.

Dedicated bearer. It is a bearer linked with a default bearer established previously that provides a dedicated tunnel to give an appropriate treatment to specific services or traffic. It does not require separate IP address and can be classified as guaranteed bit rate (GBR) and non-guaranteed bit rate (non-GBR). GBR bearer has dedicated network resources and is used for real-time voice and video applications. A non-GBR bearer does not have dedicated resources and is utilised for best-effort traffic. Dedicated bearers use Traffic Flow Templates (TFT) to provide special treatment to specific services.

QoS parameters per EPS-bearer

Each bearer uses a set of QoS parameters to describe the properties of the transport channel, such as bit rates, packet delay, packet loss, bit error rate and scheduling policy. The traffic running between a particular client application and a service can be differentiated into separate Service Data Flows (SDFs).

The four key parameters are:

QoS class indicator (QCI): The QCI specifies the treatment of IP packets received on a specific bearer (e.g. scheduling weights, admission thresholds, queue management thresholds, link-layer protocol configuration, etc.). This treatment is handled by each functional node (e.g. eNodeB, PDN-Gateway). The 3GPP has defined a series of standardised QCI types, but the operators may define proprietary QCIs to introduce new services.

Allocation and Retention Priority (ARP): The ARP is used for deciding whether new bearer modification or establishment request, connection setup and release, should be accepted considering the current resource situation. Each bearer has an associated allocation and retention priority that can be used by the eNodeB to decide which bearer(s) to drop in case of resource limitations or traffic congestion.

Guaranteed bit rate (GBR): GBR determines the bit rate that the network guarantees per EPS bearer. Specified independently for uplink and downlink. In 3GPP Release 8 and beyond, the MBR must be set equal to the GBR; that is, the guaranteed rate is also the maximum rate that is granted by the system. It is applicable only for real-time services

Maximum bit rate (MBR): MBR specifies the maximum guaranteed bit rate per EPS bearer. Specified independently for uplink and downlink. It is applicable only for real-time services.

The other important parameters associated with each bearer type are:

- APN Aggregate Maximum Bit Rate (A-AMBR): maximum non-GBR throughput allowed to specific APN. It is applicable only for non-GBR bearers.
- UE Aggregate Maximum Bit Rate (UE –AMBR): maximum non-GBR throughput allowed among all APN to a specific UE. It is applicable only for non-GBR bearers.

- Traffic Flow Template (TFT): TFT determines rules so that UE and Network are aware which IP packet should be sent on particular dedicated bearer.
- Linked EPS Bearer ID (L-EBI): L-EBI informs dedicated bearer which default bearer it is attached to.

Table 7-5: Overview bearer types and QCI classes

QCI	Bearer Type	Priority	Packet Delay	Packet Loss	Example
1	GBR	2	100 ms	10^{-2}	VoIP call
2		4	150 ms	10^{-3}	Video call
3		3	50 ms		Online Gaming (Real Time)
4		5	300 ms	10^{-6}	Video streaming
5	Non-GBR	1	100 ms		IMS Signaling
6		6	300 ms		Video, TCP based services e.g. email, chat, ftp etc
7		7	100 ms	10^{-3}	Voice, Video, Interactive gaming
8		8	300 ms	10^{-6}	Video, TCP based services e.g. email, chat, ftp etc
9		9			

7.14.2. QoS innovation in 5G Norma

The concept followed in the mobile network design so far, from the deployment point of view, is based on a limited approach where the network is deployed like a static block and the services (basically voice and data) are implemented over the network. The service is adapted to the network.

In the future 5G networks and specifically in 5G NORMA, the network concept changes, the network will be designed as a multi-service adaptive mobile network where the network resources fulfil the service requirements in a flexible and dynamic way. The network will be adapted to multiples services.

This network concept will raise new objectives and advanced QoS requirements that are important to consider. Currently, the QoS is related to the bearer model and uses a static set of QoS parameters associated with a QCI value that specifies the treatment of IP packets received on a specific bearer (Table 7-5). This approach is correct for legacy networks, but for the future 5G networks and 5G NORMA need new dynamic methods and parameters in order to satisfy the coming requirements from the new services. Another important point is that the current access network doesn't know anything about users and traffic. The strategy of resource schedule and allocation is from the QCI indication of the core network. 5G networks should have the capability of awareness and adaptation of the user and the traffic.

When imposing requirements on QoS in 5G networks, two key traffic models should be considered: high-speed video flow "server-subscriber" and massive M2M. QoS management mechanisms in 5G networks should provide video and VoIP traffic prioritization towards Web search traffic and other applications tolerant to QoS. In this line, some of the most important parameters to be taken into account are the total packet delay budget and the packet error lost rate.

5G NORMA aims to break the QCI concept used in LTE and design and implement a dynamic QoS method that provides flexibility to fit the QoS parameters to service needs and network resources and support any kind of service in a dynamic way. In order to provide such flexibility, there will be two different ways to define an EPS bearer, dynamic and static, based on the service type. Using the dynamic way, the QoS parameters will be set up on the fly, providing more responsiveness. On the other hand, using the static way, the QoS parameters will be managed beforehand, getting in this way more scalability.

The development of the NFV concept will lead to virtualization of quality management function that will be introduced in the form of two main function blocks: QoS management function block (in accordance with SLA service contracts) and the QoS control function block (real time control of traffic flows). Following this approach and the integration with the 5G NORMA architecture the main points considered are as follows: (1) Manage a flexible initial set of parameters at SLA level that will be treated by the different orchestrators and translated to specific dynamic QoS parameters at network level. (2) Manage a set of QoS parameters at network level, dynamically configurable that will be treated by the different network controllers. At this level, it would be made QoS monitoring and QoS enforcement tasks (Figure 7-47).

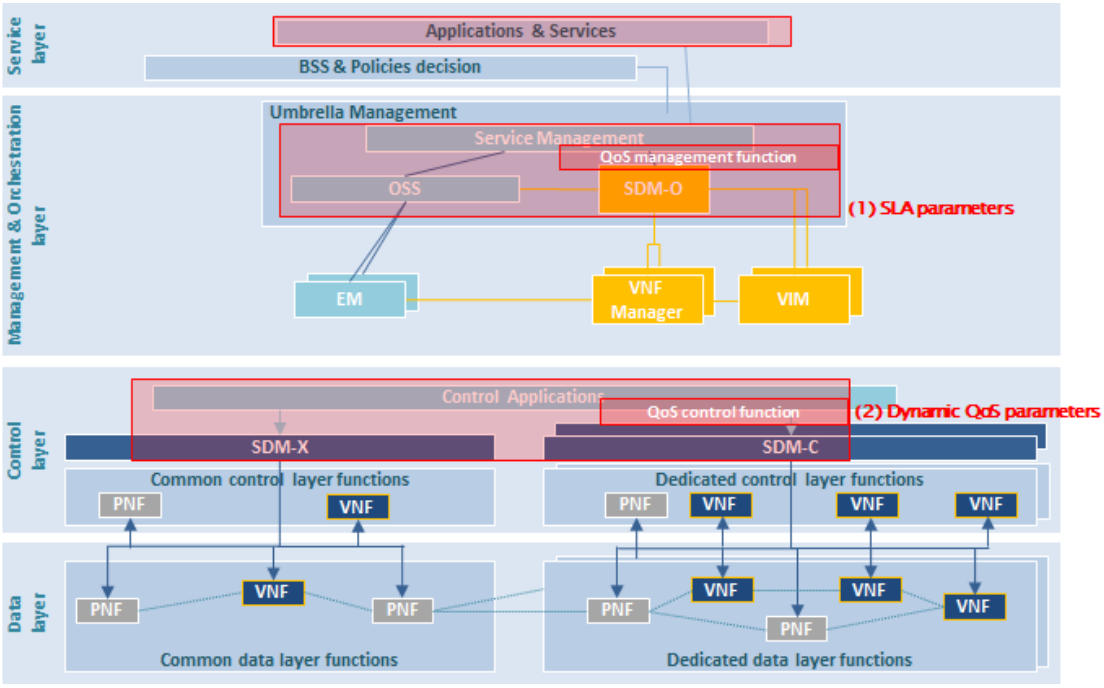


Figure 7-47: QoS innovation in 5G NORMA architecture

5G NORMA support fully dynamic, context aware quality service management able to adapt the end-to-end resource allocation and the data layer services accordingly. In order to make this, it is important to get information about the service performance and use it to adjust the relevant service parameters to try to improve it optimizing the available resources.

Monitor and control of QoS

The network is usually examined objectively by measuring a number of objective criteria in order to determine the network quality. A set of service parameters will be exposed by the applications and services executed on the 5G NORMA architecture. The QoS management function block (see Figure 7-47) will interpret these parameters at SLA level and translate them into specific dynamic QoS parameters at network level (i.e.: latency, jitter, packet loss). The SDM-O in charge of the common and dedicated functions selection (PNFs, VNFs) and with knowledge of mobile network functions (RAN domain) will take into account these parameters when the service is deployed. It is assumed that from this set of parameters it should be possible to assign the correct weight to each parameter in order to get the expected QoS.

The QoS control function block will be in charge of the network monitoring and configuration in real time through open interfaces that interact with the SDM-X, SDM-C and the control radio stack in the control layer that will integrate the specific requirements of each user. For the QoS monitoring tasks, two approaches can be taken into account: ‘intrusive’ or ‘non-intrusive’. The non-intrusive methods are purely based on monitoring the already available QoS parameters. On

the other hand, intrusive methods are based on installing specific purpose applications to get additional QoS parameters. Other solutions are focused on including new network elements (e.g., network probes and analysers, deep packet inspectors, etc.) that are responsible for capturing the traffic from a certain service and analysing its performance.

A QoS agent (specific piece of software) in network elements (VNFs, PNFs) will monitor the QoS status at run time, e.g., connection speed, packet loss rate, etc. When some monitoring parameter does not fulfil the proper values, an event will be generated and captured by the SDM-C/X and radio control functions. Then, the QoS control function block will receive the reports of these events from the controllers and will aggregate and analyse the data, enforcing new configuration actions (resource re-scheduling, new resource assignments) on the network elements (VNFs, PNFs) to meet QoS requirements of each mobile service according to the assessment results. The adjustment happens when the evaluation result is below a threshold that is defined by service requirements. If some actions are required from the management layer, QoS control function block will come back to QoS management function block in order to enforce new actions.

Figure 7-48 depicts the main interactions previously refereed:

- QoS Management – QoS Control: used to communicate the QoS Control function block with the QoS management function block.
- QoS Control – Data layer: used to receive QoS data from the network elements in the case of network monitoring, and to send configuration information in the case of network configuration.

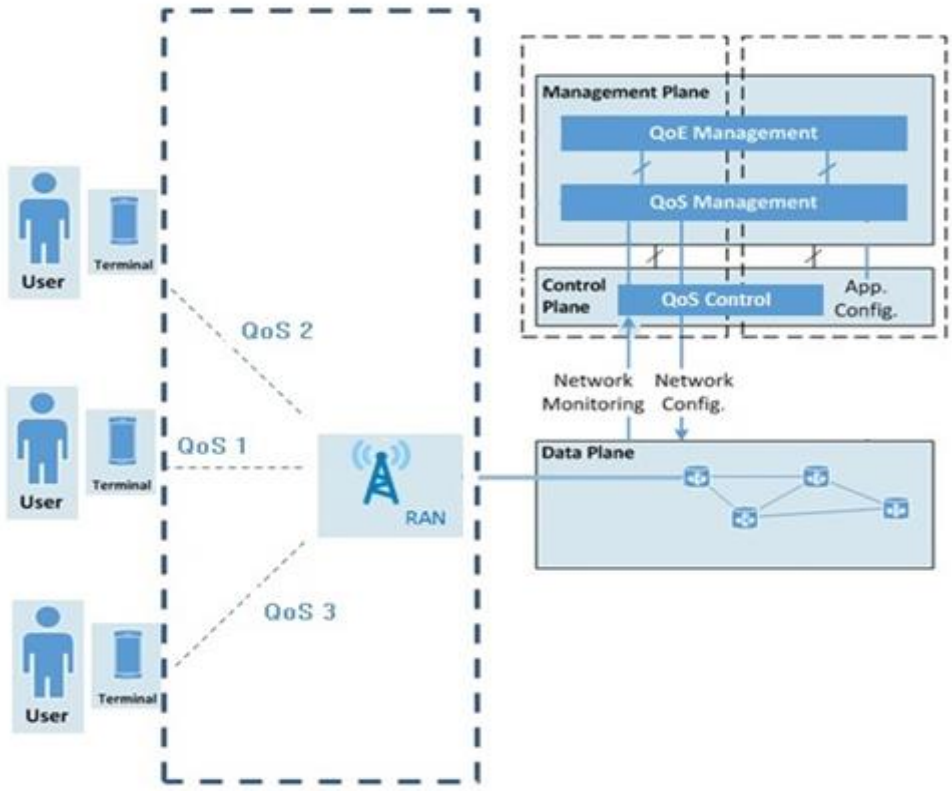


Figure 7-48: Monitor and control of QoS.

QoS metrics

We will have QoS data from different sources and expressed in a very different way. QoS data source can be diverse, e.g., radio stack parameters, profile users databases, billing information, or

values from different network interfaces, and the parameters format can be different depending on specific encoding schemas.

The QoS data set defined in each service will be named QPS (QoS Parameters Set) and it will be dynamically configurable and managed by the different controllers in the control layer. For instance:

$$\mathbf{QPS}_{\text{service a}} = \textit{Parameter 1 [0…n], Parameter 2 […], …, Parameter n […]} \quad (7.14\text{-}1)$$

Some specific parameters which are commonly used to identify service level objectives are response time, delay, jitter, data rate, required bandwidth, loss rate, error rate, round trip time (RTT), received signal strength indicator (RSSI).

8. Integration of Innovations into Functional Architecture: Process Examples

As detailed so far, the control and data layer architecture involves a novel differentiation into distributed, common (SDM-X) and dedicated (SDM-C) control functions that jointly control a distributed end-to-end data layer consisting of partly common, partly dedicated slice-individual network functions. While such a differentiation is novel by itself, the c/d-layer architecture additionally supports various 5G NORMA innovations, which are presented later, especially in Part II, of this deliverable. Here, with the help of example processes and for selected innovations, their implementation on the c/d-layer architecture will be explained.

The first set of processes in subsection 8.1 shows the implementation of the novel user-centric connection area that efficiently supports mobility with low signalling overhead. Then multi-RAT related processes are shown, first the single link case specifically useful for mm-wave integration in subsection 8.2, then the multi-link case in subsection 8.3 for increased throughput and/or robustness.

The flexibility of the decomposed control and data layer can be used to specifically adapt to the service, first by adapted configuration of the network functions themselves as exemplified in subsection 8.4, then through advanced QoS/QoE support down to the elementary flow level in subsection 8.5 and incorporating QoS monitoring in subsection 8.6. Radio resource management related processes follow, showing the use of centralized management to control TDD patterns for virtual cell creation in subsection 8.7 and dynamic resource allocation among multiple tenants in subsection 8.8.

Processes that integrate D2D group communication specifically for mMTC are given in subsection 8.9. Subsection 8.10 explains the balancing of signalling load among network function instances in the network. Finally, the process to make use of geolocation information is generally explained in subsection 8.11.

8.1. User-centric connection area

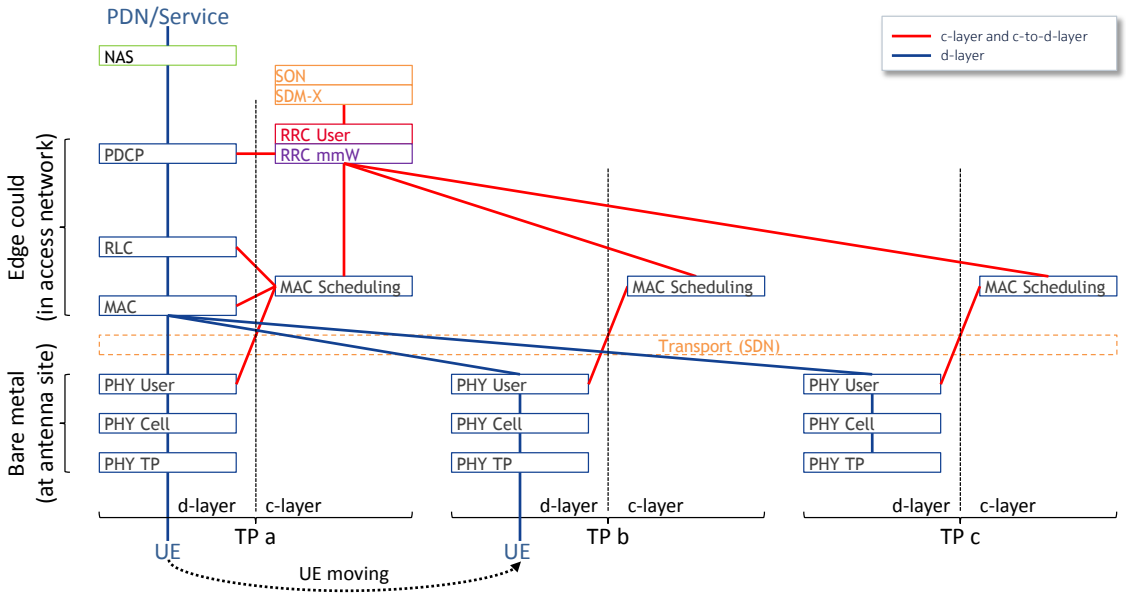


Figure 8-1: Functional c/d-layer and deployment architecture for user-centric connection area setup and best cell update

Figure 8-1 depicts the subset of functions blocks involved in respect to UCA setup and best cell update, and shows the deployment to antenna site, edge and core cloud assumed in the following process descriptions. For the complete c/d-layer architecture, refer to D4.1 Part I section 3.1.

The interactions of these function blocks are shown in the following two message sequence charts Figure 8-2 for the setup of a UCA and Figure 8-3 for the best cell update within an already setup UCA. Detailed explanations of the shown processes can be found in the associated process descriptions in Table 8-1and Table 8-2, respectively.

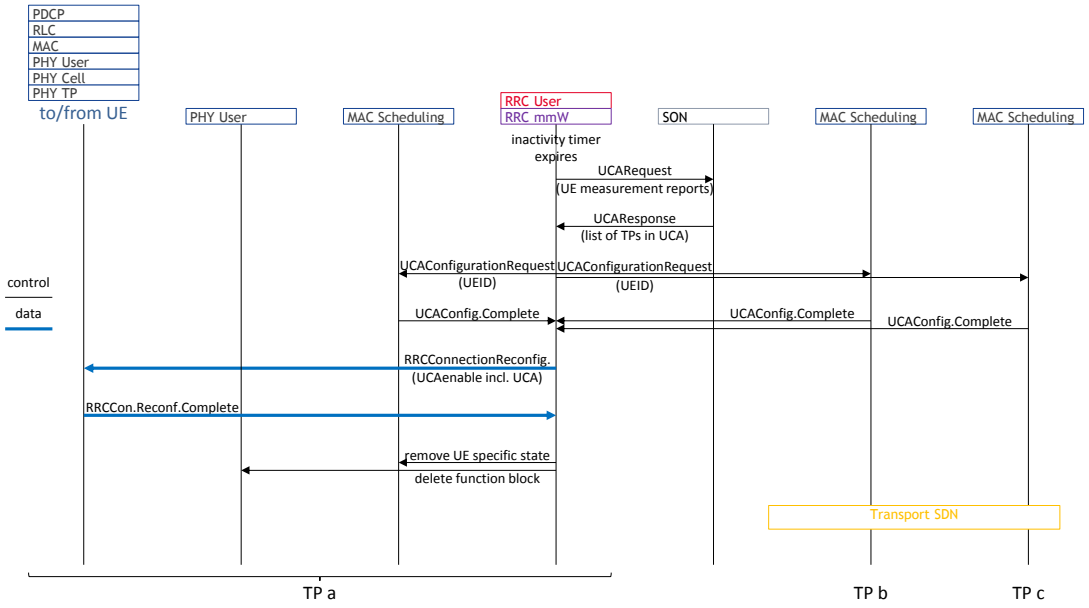


Figure 8-2: MSC user-centric connection area setup

Table 8-1: Process user-centric connection area setup

Actors:	**Triggering actor:** • RRC User/RRC mmW **Involved actors:** • SON • RRC User/RRC mmW • MAC Scheduling (RRM) • MAC, RLC, PDCP • PHY TP, PHY Cell, PHY User • Transport (SDN)
Preconditions:	• UE is UCA-capable. • UE is in RRC CONNECTED, substate UCA disabled. • *SON* knows NRT of UE's current serving cell/TP. • *RRC User* has received measurements from UE.
Postconditions:	• UE is in RRC CONNECTED, substate UCA enabled. • UE's L2 (MAC, RLC, PDCP) state from RRC CONNECTED is retained, while L1 state has been released.

	• UCA is setup: all TPs' *MAC Scheduling (RRM)* within UCA are aware of the UEID and its *MAC* instance to forward any corresponding UL packet received via Small Packet Transmit Procedure (SPTP) • Network side packet forwarding from all TPs in UCA to anchoring *MAC* is configured (can be connectionless).
Frequency of Use:	Triggered per TTI per TP; on average expected to be much less.
Normal Course of Events:	1. UE inactivity times expires at *RRC UE/RRC mmW*. 2. *RRC User/RRC mmW* requests *SON* to provide it with a suitable UCA for the UE. 3. *SON* defines UCA based on included measurement reports and applying UE-specific anticipatory (mobility prediction) selection of the UE's likely future neighbourhood, and provides a suitable UCA, i.e. a set of TPs, back to *RRC User/RRC mmW*. 4. *RRC User/RRC mmW* pushes UE context (UEID and which *MAC* instance is responsible) to all TPs' *MAC Scheduling (RRM)* within UCA, which acknowledge successfully reception of UE context (or decline e.g. to insufficient resources). 5. *RRC User/RRC mmW* sends an RRCConnectionReconfiguration message (via data layer protocol stack PDCP, RLC, MAC, PHY) to the UE to enable UCA, which the UE acknowledges with RRCConnectionReconfigurationComplete. 6. *RRC User/RRCmmW* informs *MAC Scheduling (RRM)* about the switch from RRC CONNECTED, substate UCA disabled to RRC CONNECTED, substate UCA enabled, on which *MAC Scheduling* removes UE's RRC CONNECTED state from MAC and removes PHY UE instance. 7. *RRC User/RRC mmW* triggers *SDM-C* via *Transport SDN* to configure network side packet forwarding from all TPs in UCA to anchoring MAC UE (can be connectionless configured)
Alternative Courses:	• **Distributed RAN**: Separate *RRC User/RRC mmW* instance per each TP; on reception of an UL packet via SPTP, setup of *MAC* and *RLC* instances with its state reset, i.e. any buffer status and PDCP segments a UE may have sent via another TP beforehand are lost; finally forwarding of RLC SDUs to anchoring *PDCP*. • **RRC Extant** [F5G]: while UCA employs UE context push, i.e. in advance distribution of state and possible network UL forwarding configuration, RRC Extant employs a UE context fetch, i.e. on demand request of state on reception of an UL packet via SPTP.
Exceptions:	
Assumptions:	• C-RAN with L1-L2 resp. MAC-PHY split: distributed L1 (PHY) at antenna sites and centralized L2 (MAC, RLC, PDCP) in the edge cloud (for alternatives cf. below), i.e. all UE-specific function blocks (except PHY UE) are located within the edge cloud in the access network. • UCA spans only TPs handled by this edge cloud.
Notes and Issues:	• *RRC User/RRC mmW* function block is the *RRC User* function block with included *RRC mmW* subblock, i.e. extended with novel mm-wave related functionality. • It is to be decided how to reflect a stateless L1, i.e. an L1 not maintaining any per user session, with the function block model employed by 5G NORMA; either a per single *PHY User* instance

	would be transiently created based on control information from *MAC Scheduling (RRM)* to L1 processing either alone (UL) or accompanied by a TB from MAC (DL); alternatively a single *PHY User* common to all UEs and configured with a certain type of FEC, receiver, modulation etc. could be envisioned. For the former, *PHY User* scales with the number of active UEs per TTI, for the latter with the number of different active UE classes, releases, services and/or RATs supported by that TP.

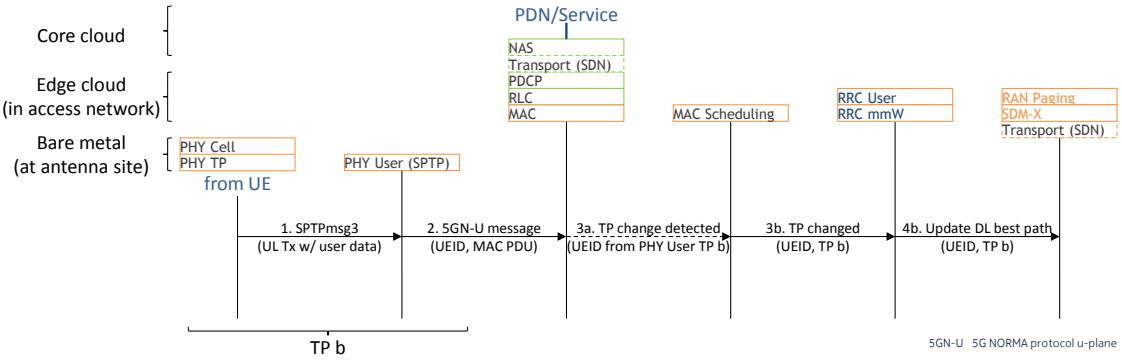


Figure 8-3: MSC user-centric connection area best cell update

Table 8-2: Process user-centric connection area best cell update

Actors:	**Triggering actor:** • PHY UE, by Small Packet Transmit Procedure (SPTP) **Involved actors:** • RRC User/RRC mmW • MAC Scheduling (RRM) • MAC, RLC, PDCP • PHY TP, PHY Cell, PHY User • Transport (SDN)
Preconditions:	• UE is in RRC CONNECTED, substate UCA enabled. • UCA is setup: all TPs' MAC Scheduling (RRM) within UCA are aware of the UEID and its MAC instance to forward any corresponding UL packet received via SPTP. • Network side packet forwarding from all TPs in UCA to anchoring MAC is configured; can be connectionless configured by Transport (SDN).
Postconditions:	• UE is in RRC CONNECTED, substate UCA enabled. • Network side packet forwarding of DL packet is updated to TP to which the UE has moved and sent its last UL packet.
Frequency of Use:	Triggered per TTI per TP; on average expected to be much less.
Normal Course of Events:	1. An UL transmission is received by via SPTP (involving *PHY TP*, *PHY Cell*, *PHY User* (SPTP). 2. *PHY User* (SPTP) uses the UEID contained in the UL transmission to determine the UE's *MAC* instance to which it forwards the received MAC PDU. 3. *MAC Scheduling (RRM)* detects that MAC PDU is received from a different TP's *PHY UE* than before and indicates the new UE location to *RRC UE*.

	4. *RRC User/RRC mmW* triggers *SDM-C* via *Transport (SDN)* to update DL forwarding state to establish the last TP as the best cell for a possible future DL packet to this UE.
Alternative Courses:	• **SPTP with common "lower" MAC**: UEID is included as MAC control element, therefore a *"lower" common MAC* for processing only SPTP-specific parts is inserted in between *PHY User* (SPTP) and user-specific *MAC*; the latter is then selected based on the UEID and takes over processing the remainder of the MAC PDU incl. e.g. SDU demux. • **SPTP with user-specific "upper" PHY User**: UEID is included in PHY header of SPTPmsg3, which is processed by the "lower" *PHY User* (SPTP), is inserted in between *PHY Cell* and user-specific *PHY User*; the latter is then selected based on the UEID and takes over processing of the remainder of SPTPmsg3.
Exceptions:	
Assumptions:	• C-RAN with L1-L2 resp. MAC-PHY split: distributed L1 (PHY) at antenna sites and centralized L2 (MAC, RLC, PDCP) in the edge cloud (for alternatives cf. below), i.e. all UE-specific function blocks (except *PHY User*) are located within the edge cloud in the access network. • UCA spans only TPs handled by this edge cloud.
Notes and Issues:	• *MAC* may be split into two function blocks: Scheduling UEs in RRC CONNECTED mode and SPTP scheduling. • Involvement of SDM-C function block with respect to *Transport (SDN)* needs to be further specified.

8.2. Multi-RAT integration

In the following the required functions blocks and their interaction are shown in Figure 8-4 for the setup of a mm-wave multi connectivity and in Figure 8-5 for a UE based (forward) handover within the mm-wave architecture.

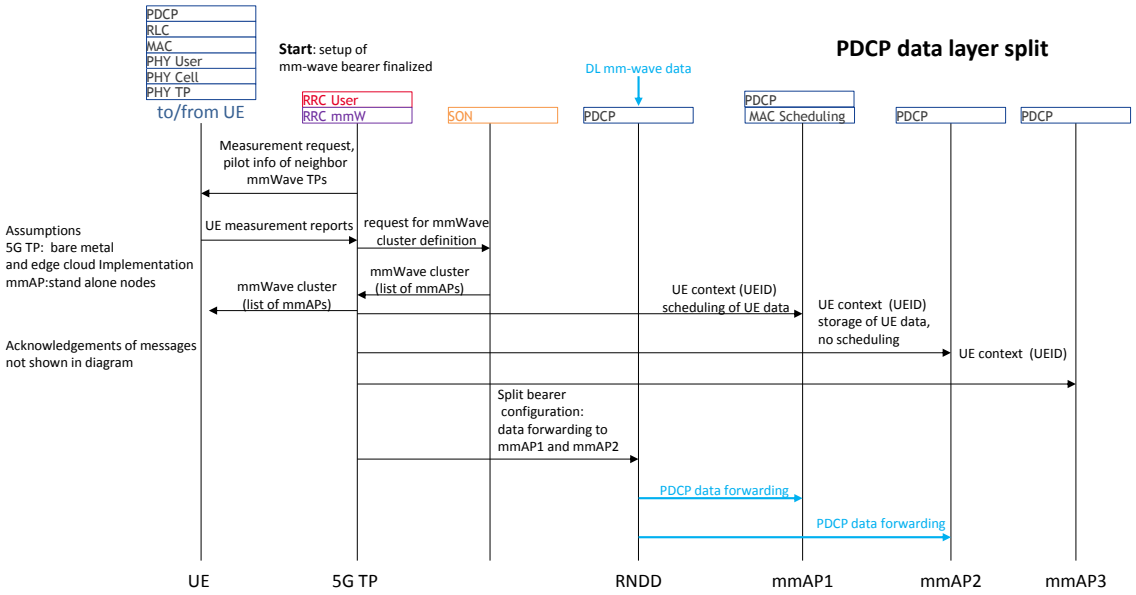


Figure 8-4: MSC setup of mm-wave multi connectivity

In the following, the setup of the mm-wave multi connectivity process is described.

Table 8-3: Process setup of mm-wave multi connectivity

Actors:	**Triggering actor:** • RRC User/RRC mmW **Involved actors:** • PDCP • RRC User/RRC mmW • SON
Preconditions:	• Setup of mm-wave bearer is finalized. • SON knows NRT of UE's current serving cell/TP.
Postconditions:	• Mm-wave cluster defined by SON. • Mm-wave cluster communicated to UE. • mmAPs of defined cluster prepared.
Normal Course of Events:	1. *RRC User* requests UE to carry out measurements of surrounding mm-wave TPs, for which it provides the UE with pilot information of neighbour mm-wave TPs. 2. UE reports mm-wave TPs which could be measured. 3. *RRC mmW* requests *SON* to defined a cluster of mmAP for UE. 4. *SON* defines based on NRT and anticipatory information the mmAP cluster for the UE and transfers the info to *RRC mmW*. 5. *RRC User* informs the UE about its mmAP cluster. 6. *RRC mmW* provides all mmAPs of the cluster with the UE context. 7. *RRC mmW* informs PDCP (node PDCP H) about APs (nodes) to which the PDCP data should be forwarded. Information about buffering only or buffering and scheduling is included.
Special Requirements:	• UE can only measure surrounding mmAPs if it has been provided with pilot information of surrounding mmAP.
Assumptions:	• Neighbour list of surrounding mmAPs is provided by SON to *RRC User* beforehand of the presented process, e.g. during setup of the 5G TP.
Notes and Issues:	• If the UE moves to a border of a mm-wave cluster, pilot information of mmAPs not belonging to the cluster have to be provided to the UE to enable a handover to mmAPs not belonging to the cluster.

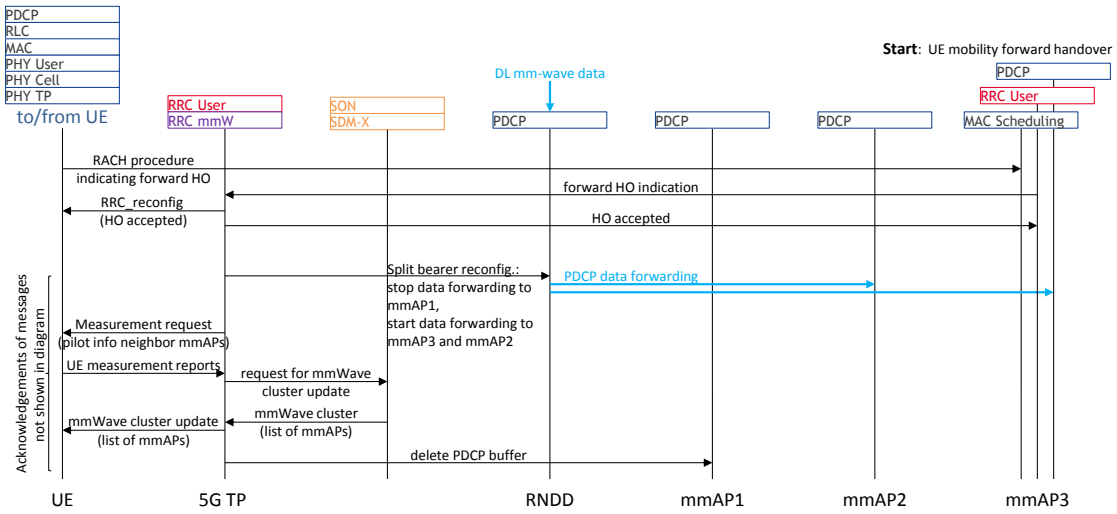


Figure 8-5: UE mobility within mm-wave architecture for multi connectivity

In the following, the mobility within mm-wave multi connectivity process is described.

Table 8-4: Process UE mobility within mm-wave architecture for multi connectivity

Actors:	**Triggering actor:** • RRC User **Involved actors:** • MAC Scheduling • PDCP • RRC User/RRC mmW
Preconditions:	• UE is in RRC CONNECTED mode. • Cluster of mm-wave APs is defined. • UE is served by mmAP1.
Postconditions:	• UE is served by mmAP3. • mmAP1 is no longer involved in transmission of mm-wave data.
Normal Course of Events:	1. UE moves towards mmAP3 and detects pilots of mmAP3 better than pilots of serving mmAP1. 2. UE starts RACH procedure (*MAC Scheduling (RRM)*) indicating forward HO to mmAP3. 3. *RRC User* (mmAP3) informs *RRC User* of 5G TP about HO indication and carries out admission control etc. 4. *RRC User* (5G TP) reconfigures UE, indicating that the HO is accepted. 5. *RRC User* (5G TP) informs *RRC User* (mmAP3) that HO is accepted. 6. *RRC User* (5G TP) informs *PDCP* (node PDCP_H) to stop forwarding PDCP data to mmAP1 and to forward PDCP data to mmAP3. 7. *PDCP* (node PDCP_H) starts forwarding PDCP data to mmAP3. 8. *RRC User* (5G TP) User request UE to start measurements to define a new mmAP cluster (details see above process for setting up mm-wave multi-connectivity). 9. RRC User (5G TP) requests *PDCP* of mmAP1 to delete PDCP data as mmAP1 no longer belongs to UE specific mmAP cluster.
Special Requirements:	• UE is controlled during the complete HO process by 5G TP to enable a secure exchange of control messages.
Assumptions:	• UE is allowed to carry out a forward handover, which reduces delay caused by the HO compared to a backward handover.
Notes and Issues:	• Forward handover is currently not accepted by 3GPP for LTE.

8.3. Multi-connectivity support in multi-RAT networks

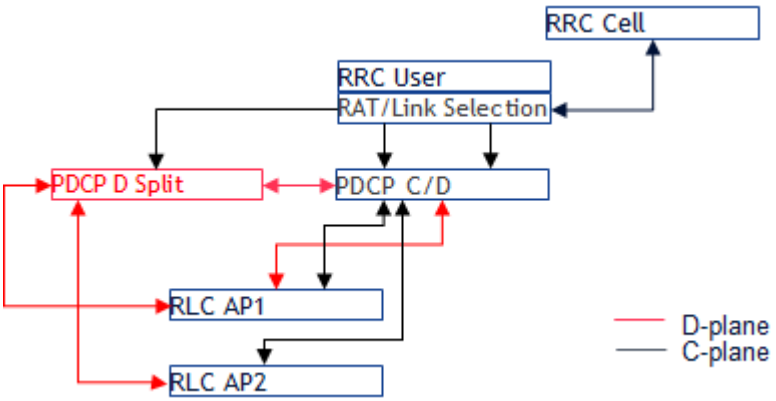


Figure 8-6: C/d-layer architecture for multi-connectivity support in multi RAT environment

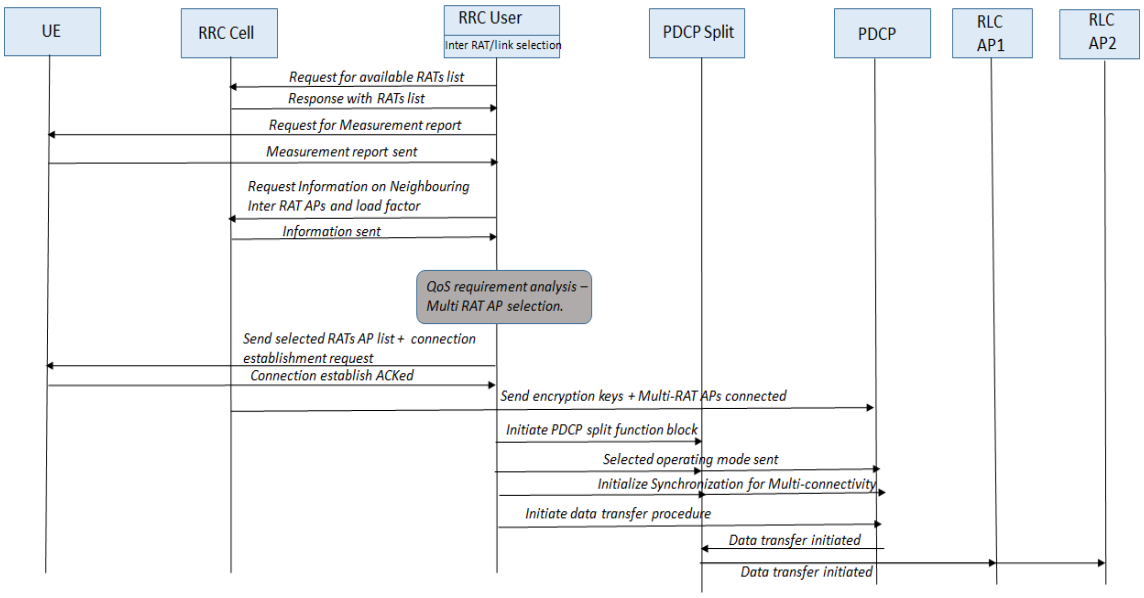


Figure 8-7: MSC for multi-connectivity support in Multi-RAT networks

Table 8-5: Process for multi-connectivity support in Multi-RAT networks

Actors:	**Triggering actor:** • Inter RAT/Link Selection as a part of RRC User Involved actors: • PDCP • PDCP Split Bearer • RRC User
Preconditions:	• UE connects to two or more RATs simultaneously. • Common PDCP and RRC layer
Postconditions:	• Best Inter RAT link and data transfer operating mode selected • Efficient data transfer to UE using multiple RATs simultaneously • Quality of Service based operating mode selected

Normal Course of Events:	1. RRC Cell for the list of the available RATs to which UE can connect. RRC Cell provides the information of all the neighbouring cells and user preferences if predefined. 2. RRC User function requests UE the measurement report about the Radio Signal Conditions from the set of APs, User preferences etc. 3. The UE sends the measurement report to RRC user that can be later used for multi-connectivity connections establishment. RRC user now has the requested measurement report for a set of APs that may belong to different RAT. The measurement information can be categorise into both, Intra RAT as well as Inter RAT. 4. RRC User function block select the RAT by mapping it on to the QoS requirements of UE. 5. Depending on the Inter RAT selection algorithm a decision about most feasible RATs connection to UE is made. a) The algorithm selects the inter RAT link by considering the RAT load conditions, User's preferences towards particular RAT, RAN parameters etc. b) E.g., a modified proportional fair algorithm can be implemented to manage the resources across the different RATs. 6. UE connection is then establish according to the selected RAT AP list. 7. RRC Cell transfers the encryption keys to the PDCP block. 8. PDCP Split Bearer function block is instantiated by RRC User (Inter-RAT/Link Selection). 9. The Inter RAT/link selection block in the RRC User then selects the operating mode either reliability mode that duplicates the data over multiple connections or diversity mode that splits the data over multiple connections. Depending on the load condition and QoS requirements, the data is either duplicated over the split bearer or multiplexed that is similar to the concept of dual connectivity in LTE. 10. As per the operating mode selected, initiate the packet data synchronization on PDCP and PDCP Split Bearer functional block. 11. Start the data transfer procedure in the PDCP and PDCP Split Bearer functional block. 12. PDCP Split Bearer transfers data to the RLC of connected APs as shown in the figure.
Special Requirements:	• UE support for multi-connectivity
Assumptions:	• Common RRC and PDCP for Multi-RAT. RRC PDCP located in the edge cloud.
Notes and Issues:	• South bound interface between Inter RAT/link selection and PDCP Split Bearer and PDCP functional block. Also in case of operating mode selection: data duplication or data reliable mode. If reliable mode is selected, i.e. the packets from same flow are multiplexed over two APs that belong to different RATs, advanced synchronisation will be required at receiver due to different lower layer protocols, format and transmission technology and scheduling.

8.4. RAN support for optimised on-demand adaptive network functions and services

Figure 8-8 shows the process with interaction between most relevant function blocks for some flexible on-demand configurations of RAN protocols' functions and services, as part of the innovation detailed in Section 2.2.1.

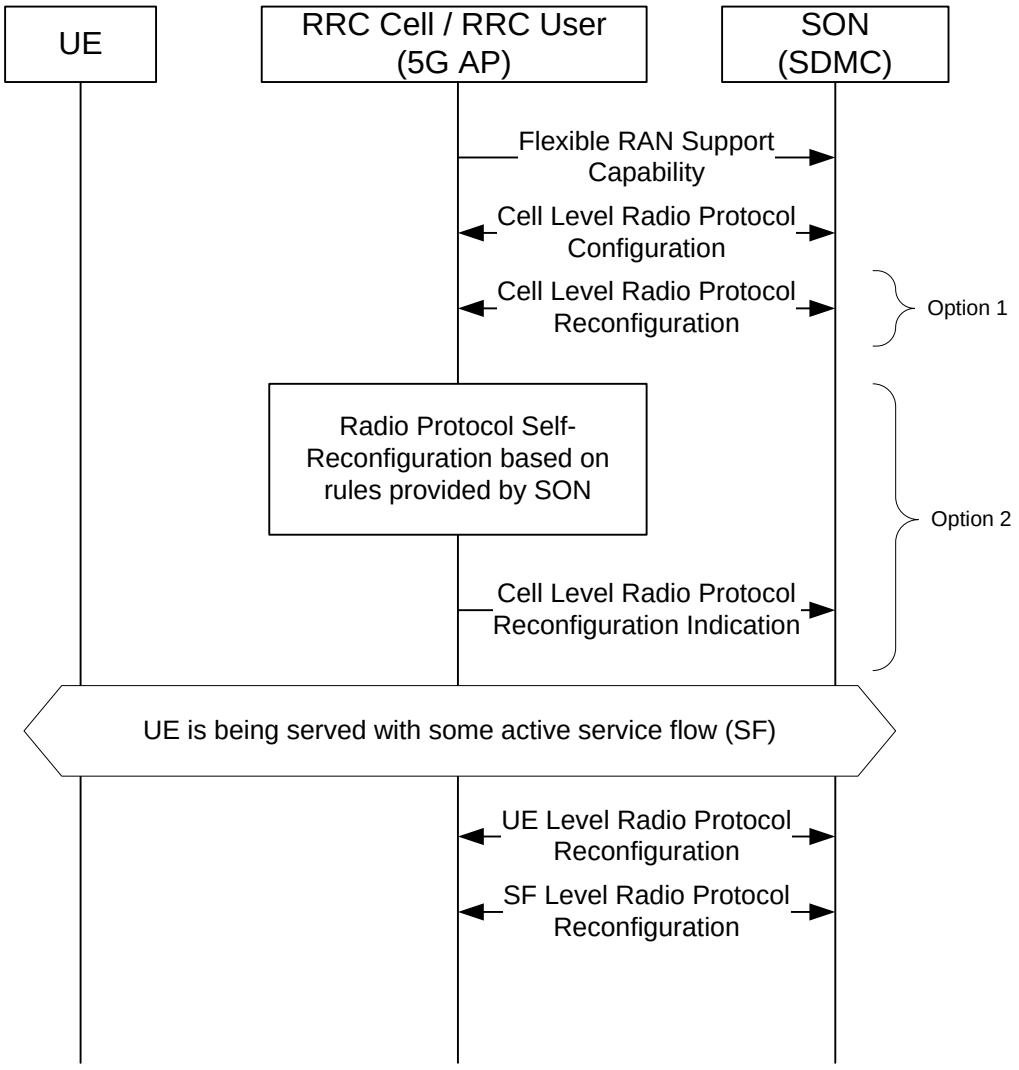


Figure 8-8: MSC for facilitating flexible radio protocol configuration

Table 8-6: Process for facilitating flexible radio protocol configuration

Actors:	Triggering actor: • RRC Cell or RRC User (at 5G AP) Involved actors: • SON (on top SDMC) • RRC Cell or RRC User
Preconditions:	• 5G AP's initial activation or reactivation with capability exposed to 5G network controller or orchestrator • UE being served for some particular service request
Postconditions:	• Optimised on-demand configuration of the serving cell and radio protocol stack's functions and services for the serving AP and/or individual UE or service flow, adapted to capability of 5G AP
Normal Course of Events:	1. *RRC Cell* indicates capability of 5G AP (and fronthaul/backhaul interfaces) to the network controller, *SON* on top of SDMC. 2. *SON* (SDMC) configures and later reconfigures *RRC Cell* with initial cell level configurations and rules for flexible on-demand configuration of RAN functions and services based upon indicated/exposed capability of 5G AP.

	3. Either *SON* (SDMC) or 5G AP's *RRC Cell* determines and configures 5G AP an optimised configurations of radio protocol stack. 4. Further on-the-fly reconfigurations of the radio protocol stack's functions and services of the 5G AP may be triggered by *RRC User* following a connection or service request from an individual UE in order to best serve the request as well as the overall cell. Details can be found in Section 2.2.1.
Special Requirements:	• 5G AP (and network interfaces) capable of implementing flexible on-demand configuration of RAN functions and services.
Assumptions:	• 5G AP is implemented using advanced generic or general-purpose hardware/software platforms.
Notes and Issues:	• The process may be adaptive on the fly and on-demand to best serve the cell as well as individual UEs.

8.5. Flexible 5G service-flow (SF) with in-SF QoS differentiation and multi-connectivity

Figure 8-9 shows an example of a process with interaction between the most relevant function blocks for some RAN support for advanced QoS/QoE control, as part of the innovation detailed in Section 2.8.2.

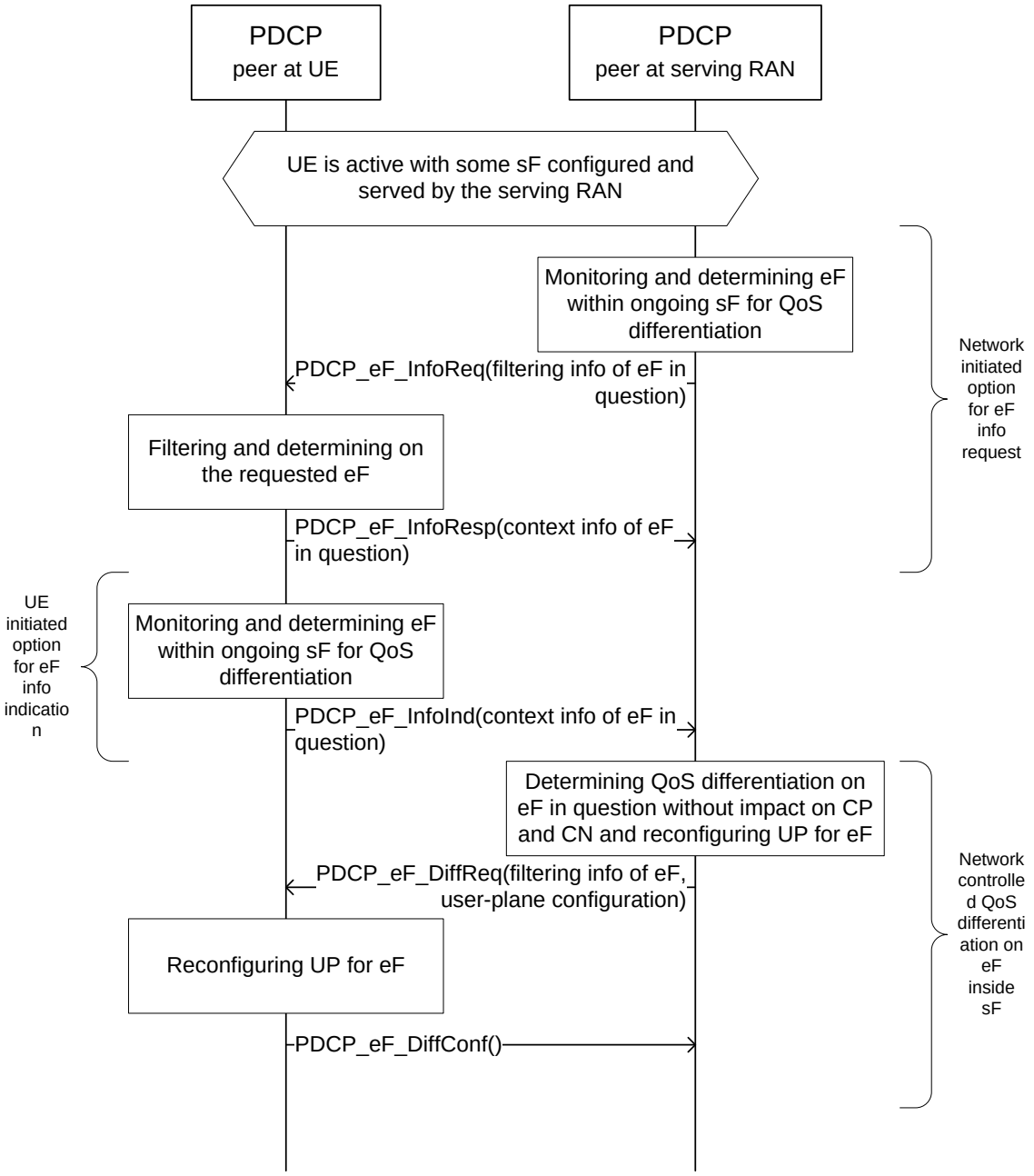


Figure 8-9: MSC for RAN support for advanced QoS/QoE control

Table 8-7: Process for RAN support for advanced QoS/QoE control

Actors:	**Triggering actor:** • PDCP **Involved actors:** • PDCP peers at UE and serving RAN
Preconditions:	• UE is configured with service- and application-aware QoS control functions for supporting in-bearer QoS differentiation. • UE is active and has ongoing service session.
Postconditions:	• Optimised RB configurations and RB services for the UE with in-RB QoS differentiation applied to some elementary service flow, referred to as eF in the MSC above, within an established RB of the ongoing service session.

Normal Course of Events:	1. PDCP peer at the UE side, triggered by either a request from *PDCP* at the serving RAN or an outcome of QoS monitoring event of *PDCP* at the UE side itself, indicates specified information on an filtered-out elementary flow (eF) such as application information, expected remaining traffic volume or session life-time, end-to-end transport protocol related information, corresponding feedback eF related information (TCP/IP-based applications), etc. 2. (a) *PDCP* at the serving RAN, based on the received eF information from the UE's PDCP, configures some optimised QoS treatments, as described in Section 2.8.2 with specific examples under: **QoS differentiation treatments in RAN level and not requiring control-data layer and RAN-CN interactions.**
Special Requirements:	• *PDCP* is able to monitor and determine on individual eFs (and corresponding eFs in other direction) of UE which can be filtered out and selected for QoS differentiation treatment. The monitoring and decision on eF may be based on, e.g., at least one of packet filtering attributes (e.g., packets' header information {source/destination IP addresses, source/destination port addresses, flow label, DSCP}, progress on the filtered eF (e.g., throughput-delay and lifetime related), etc.
Assumptions:	• *PDCP* is configured by the network controller (SDMC based *QoS Control*) with necessary rules or enforced policies for in-bearer or in-SF flow filtering and QoS differentiation.
Notes and Issues:	• The process may be adaptive on the fly and on-demand to best serve the cell as well as individual UEs.

8.6. QoS innovation

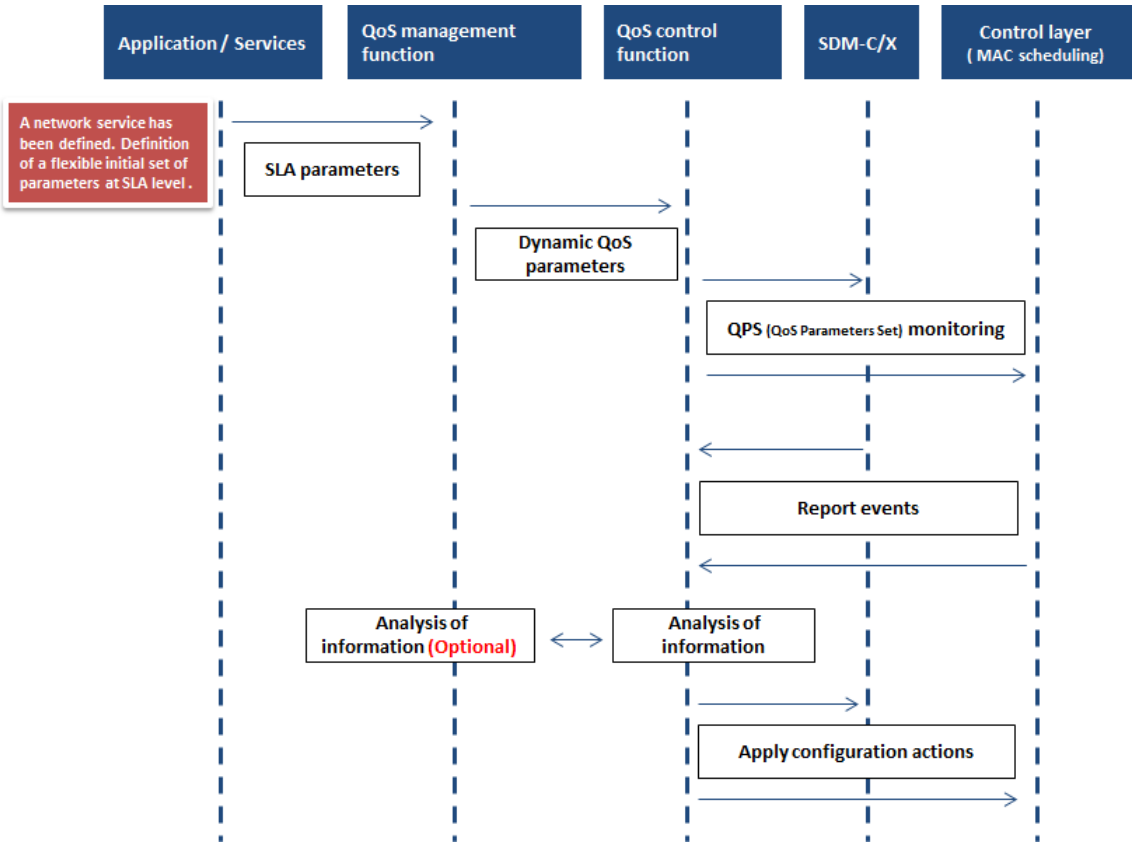


Figure 8-10: MSC for QoS innovation

Table 8-8: Process for QoS innovation

Actors:	**Triggering actor:** • QoS control function block • **Involved actors:** • • MAC Scheduling • SDM-C, SDM-X • QoS management function block
Preconditions:	• A network service has been defined. • Definition of a flexible initial set of parameters at SLA level
Postconditions:	• Continuous monitoring of the main control parameters defined for a service
Normal Course of Events:	1. A network service has been defined. Service requirements, service level agreements (SLAs) for a defined service are managed by the service management function, QoS management function block and the SDM-O when a service is deployed. 2. QoS management function manages the SLA parameters and translates these parameters to dynamic QoS parameters. 3. QoS management function block sends the dynamic QoS parameters to QoS control function block. 4. QoS control function block manages these parameters and sends to SDM-C/X and radio control blocks the QPS (QoS Parameters Set). 5. An UE attaches to the network, requiring a specific network service.

	6. Dynamic bearer activation (Dedicated bearer activation in combination with the default bearer activation) as part of the attach procedure. 7. A QPS (QoS Parameters Set) is monitored in order to fulfil the service requirements. 8. SDM-X/C and the control radio stack are monitoring the QoS parameters associated (QPS). 9. When some of the QoS parameters are not fulfilling the proper values, an event report is sent by the SDM-C/X or radio control blocks to QoS control function block. 10. QoS control function block in the control layer provides real-time control of traffic flows on the basis of QoS levels and parameters established during the connection. 11. QoS control function block applies configuration actions through the SDM-C, SDM-X and radio control blocks, when the evaluation result is below a threshold that is defined by service specifications.
Alternative Courses:	• If some actions are required from the management layer, QoS control function block will come back to QoS management function block in order to enforce new actions.
Notes and Issues:	• Basic QoS control mechanisms include traffic profiling, planning and management of data flows. • QoS management function block provides (SDMO) QoS support in accordance with SLA service contracts, as well as provides maintenance, review and scaling of QoS.

8.7. Centralised radio resource management

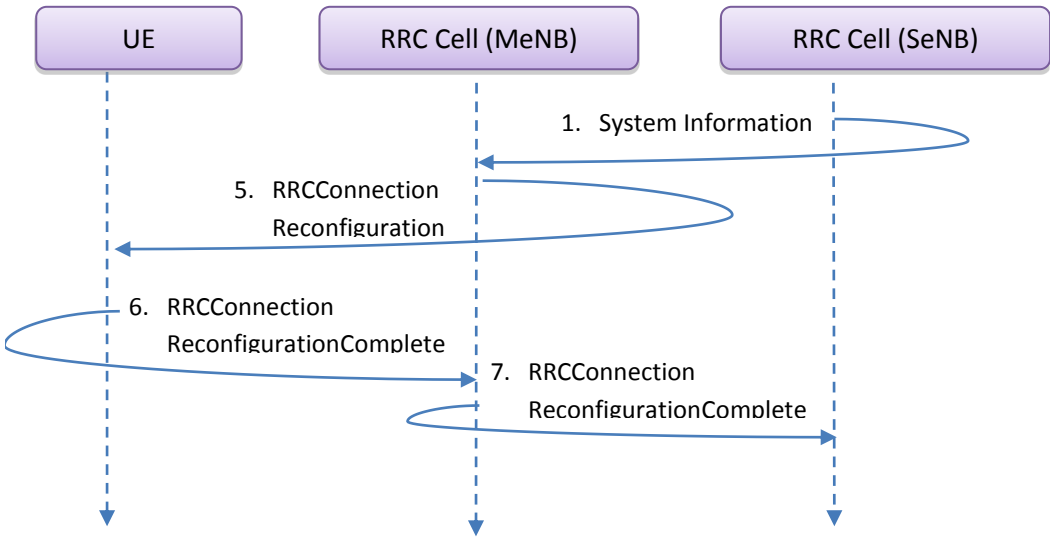


Figure 8-11: MSC for centralised radio resource management

This process describes centralised radio resource management for TDD-based networks. The innovation presented in this frame work is radio resource management for virtual cells. This process comprises

- TDD patterns for the cell(s),

- the ratio between uplink and downlink resources, and
- allocation of the resource and scheduling.

Table 8-9: Process for centralised radio resource management

Actors:	**Triggering actor:** • RRC Cell **Involved actors:** • RRC User • RRC Cell • MAC Scheduling
Preconditions:	• Each cell can have a different TDD pattern. • Cells are not clustered. • UEs keep and update a list of cells and their RSRP.
Postconditions:	• The cells are clustered based on their traffic demands. • For each cluster, the proper TDD pattern is selected. • The congested cells are defined. • Virtual cell(s) are formed to address the congestion cells, • Proper TDD-pattern for virtual cells.
Frequency of Use:	Once per several minutes to several hours
Normal Course of Events:	1. The System Information (SI) of cells in the Secondary eNodeB (SeNB) is transmitted to Master eNodeB (MeNB). 2. Congestied cells are determined. 3. Virtual cells are formed. 4. TDD pattern for each cell is determined. 5. RRCConnectionReconfiguration, including the SIs of cells SeNB, broadcasted for UEs. 6. RRCConnectionReconfigurationComplete is sent from UEs to MeNB. 7. RRCConnectionReconfigurationComplete is forwarded from MeNB to SeNBs. 8. MAC Scheduling UEs in every cell.
Alternative Courses:	1. The System Information (SI) of cells is transmitted to SDMC. 2. Congested cells are determined. 3. Virtual cells are formed. 4. TDD pattern for each cell is determined. 5. Communicate the new configuration to RRC Cells. 6. RRC Connection Reconfiguration, broadcasted for UEs, 7. RRC Connection Reconfiguration complete is sent from UEs to Me-NodeB, 8. RRC Connection Reconfiguration Complete from MeNodeB and SeNodeB, 9. MAC Scheduling UEs in every cell.

8.8. Multi-tenant dynamic resource allocation

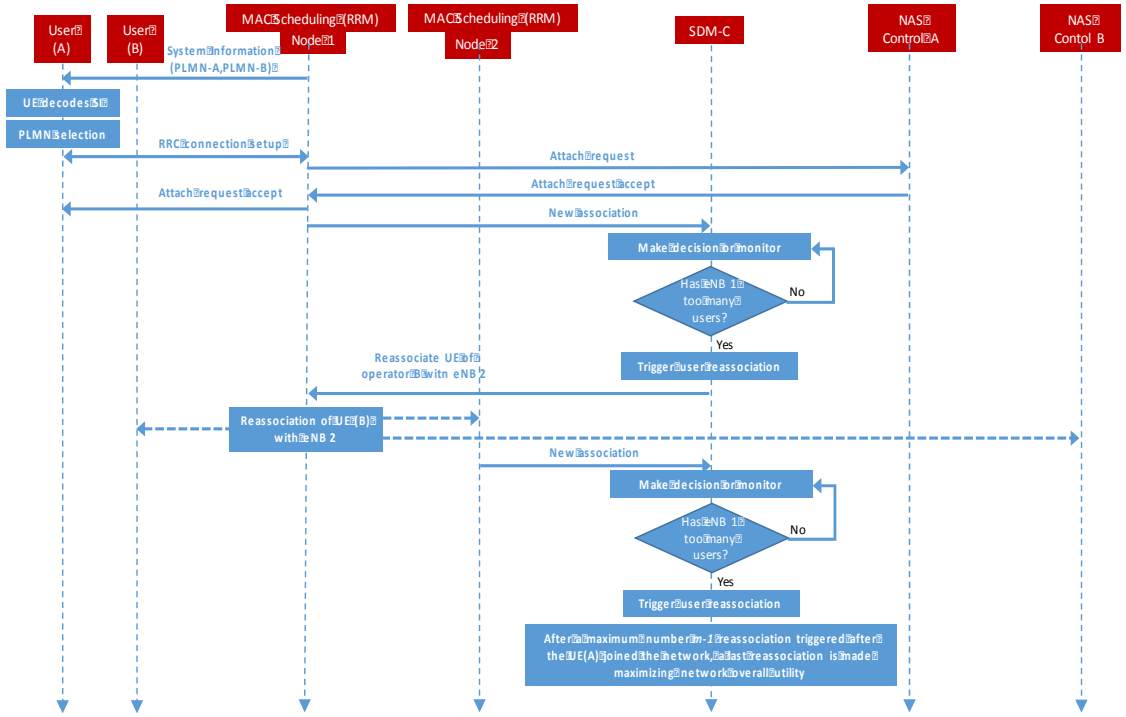


Figure 8-12: MSC for multi-tenant dynamic new user association

Table 8-10: Process for multi-tenant dynamic new user association

Actors:	**Triggering actor:** • PHY User **Involved actors:** • RRC Cell • RRC UE • NAS Control • SDM-C, SDM-X • Multi-tenancy Scheduling • MAC Scheduling (RRM)
Preconditions:	• Each operator has a certain amount of network resources accordingly with his network share s_o. • The users of a certain operator in a certain cell share the resources fairly. • The NAS Control knows the actual users association. • Multi-tenancy Scheduling sends through SDM-X the different s_o to each MAC Scheduling (RRM)
Postconditions:	• The users are associated in order to maximize network utility and share the resources fairly among UEs of each operator.
Frequency of Use:	Once a user joins the network, or it changes its location, or it leaves the network.
Normal Course of Events:	1. User receives system information from NAS Control containing the operators available (PLMN). 2. User selects the operator and starts the connection setup.

	3. The MAC Scheduling (RRM) locates to the user, resources according with the operator network share and the number of users belonging the same operator connected with the same cell. The information regarding operator network share are given by Multi-tenancy Scheduling through SDM-X. 4. If, the node 1 has too many users, the NAS Control triggers a user reassociation. 5. One of the users of node 1 is associated with another node choosing the one that allocates to the user the higher rate. 6. If, node 1 and/or node 2 now have too many users the *NAS Control* triggers the reassociation considering as candidates the users from the two nodes. 7. Eventually repeat step 6 considering users from the two nodes. 8. To avoid that the reassociation of a user harms the overall performance after a maximum number of reassociation *m-1*, the *NAS Control* associates a user with the node that maximizes the overall network utility and stops the procedure. 9. If a user leaves the network the algorithm is quite similar. 10. When a user moves within the cell, the throughput it could receive from a neighbour node changes. If the user would receive a higher throughput from the new node, the NAS Control reassociates it to this node. 11. Then, the old node executes the same algorithm as when a user leaves the network while the new node executes the algorithm corresponding to a joining user.

8.9. Multi-service technologies/mMTC

For massive machine type communication, the legacy random access procedure capacity may pose a bottle neck. This may be mitigated through forming groups of mMTC devices, for which a group controller is selected that carries out the random access procedure towards the RAN and uses device-to-device communication towards its group members. The following three processes detail first the formation of such groups (and then show how UEs can join and leave them.

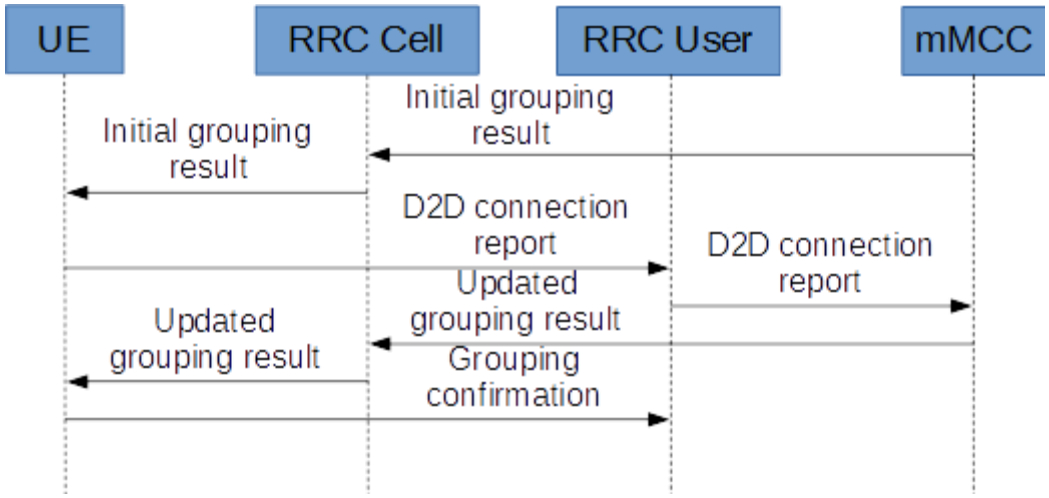
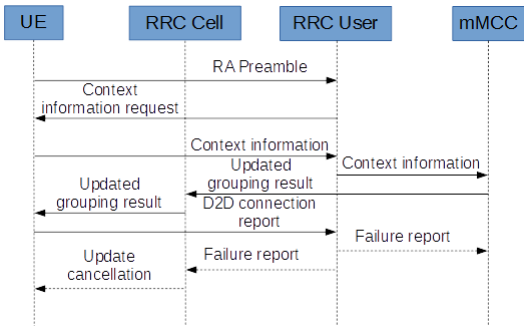


Figure 8-13: MSC for global mMTC group updating

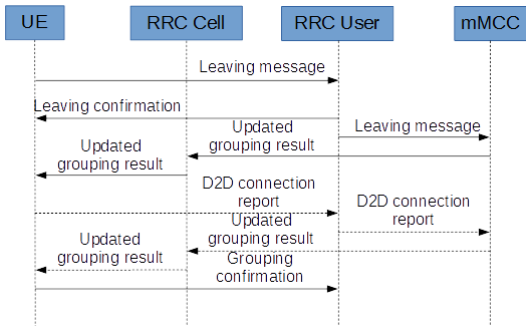
Table 8-11: Process for global mMTC group updating

Actors:	**Triggering actor:** • mMTC Congestion Control (mMCC)

	Involved actors: • RRC Cell • RRC User
Preconditions:	• The RAN level congestion rate exceeds a predefined threshold, or a timeout has passed since the last global update of groups.
Postconditions:	• The UEs are clustered into groups, for each group a group controller (GC) is selected. • Instead of directly communicating to the eNB in RACH, the UEs other than the GC work in a special group member mode that they o send their RACH requests to the GC, and receive the RACH confirmations from it, via D2D connections; o after confirmed, set up the data links directly to the eNB, as usual (or: communicate with the eNB through the GC) • The GC works in a special GC mode that it o periodically sends randomly selected preamble to eNB, to request random access for the group; o reattempts to send the request by the next opportunity, if the eNB does not response, until the request is confirmed; o helps to coordinate the time-scheduling in its group.
Frequency of Use:	Once per hours/days/…
Normal Course of Events:	1. The process is triggered by one of the preconditions, which is detected by the *mMCC*. 2. According to context information, *mMCC* clusters the UEs into groups, for each group it selects one GC and designs a time schedule. GCs and group members are informed about the result. (Note: we consider that the context information is available at the eNB, because it is uploaded during the normal RRC connection setup/reconfiguration and periodically updated.) 3. Every UE tries to access its GC via D2D connection, if it fails several times, it switches back to the ungrouped mode. 4. Every GC sends report to the *RRC User* about the successes/failures and the average D2D connection quality (throughput, SNR, etc.). 5. The *RRC Cell* informs all UEs of the grouping result. If failures occur in some groups, the *mMCC* updates their time schedules and the *RRC Cell* informs them about the update in this message. 6. The UEs confirm the grouping result and switch to (updated) group-mode.



(a)



(b)

Figure 8-14: Process for mMTC group (a) joining and (b) leaving

Table 8-12: Process for mMTC group joining

Actors:	**Triggering actor:** • RRC User **Involved actors:** • mMCC • RRC Cell • RRC User
Preconditions:	• Groups already exist in the local cell. • One UE (in synchronized mode) joins the cell.
Postconditions:	• The UE is added to an appropriate group or remains in normal un-grouped mode.
Frequency of Use:	Depends on the mobility/dynamics of UEs.
Normal Course of Events:	1. *RRC User* receives preamble sent by a UE 2. Instead of sending a normal RAN response, the *RRC User* asks the UE for its context information, including CSI, geolocation and device type. 3. The UE reports the *RRC User* about its context information 4. If the device is asynchronous, normal RAN procedure continues. If it is synchronous, the *mMCC* selects the best existing group for it (if there is any available), and draft a new time schedule for the group. 5. If a group is selected, the *RRC Cell* informs the group and the new UE about the grouping result and the new schedule. 6. The new UE tries to access its GC via D2D connection, if it fails several times, it switches back to the ungrouped mode. 7. If the D2D connection succeeds, GC reports the *RRC User* to confirm the new schedule and update the average connection quality (ACQ). 8. The involved UEs switch to updated group mode.
Alternative Courses:	1. If the *mMCC* fails to find any group for the new UE in step 4, it goes further with the normal random access procedure. 2. If the D2D connection in steps 6-7 fails, GC reports the *RRC User* to deny the new schedule and update the ACQ. *mMCC* cancels the new time schedule, records the failure, updates the ACQ, *RRC Cell* informs the involved UEs about the denial of new time schedule.

Table 8-13: Process for mMTC group leaving

Actors:	**Triggering actor:** • RRC User **Involved actors:** • mMCC • RRC Cell
Preconditions:	• Groups already exist in the local cell. • One grouped UE leaves its cell.
Postconditions:	• The involved group is updated in members and time schedule.
Frequency of Use:	Depends on the mobility/dynamics of UEs.
Normal Course of Events:	1. The UE sends leaving message to the *RRC User*.

	2. If it is a GC, the *mMCC* reselects a GC in its current group. The *mMCC* updates the time schedule of its group. 3. The *RRC User* confirms the leaving UE's message 4. The *RRC Cell* informs the involved UEs about the new time schedule (and eventually the new GC) 5. All UEs confirm *RRC User* the new time schedule and switch to the new time schedule.
Alternative Courses:	1. In case of new GC, instead of step 5: as steps 3-6 in above global mMCC group updating process, but only for the involved group.

8.10. Load balancing of signalling traffic

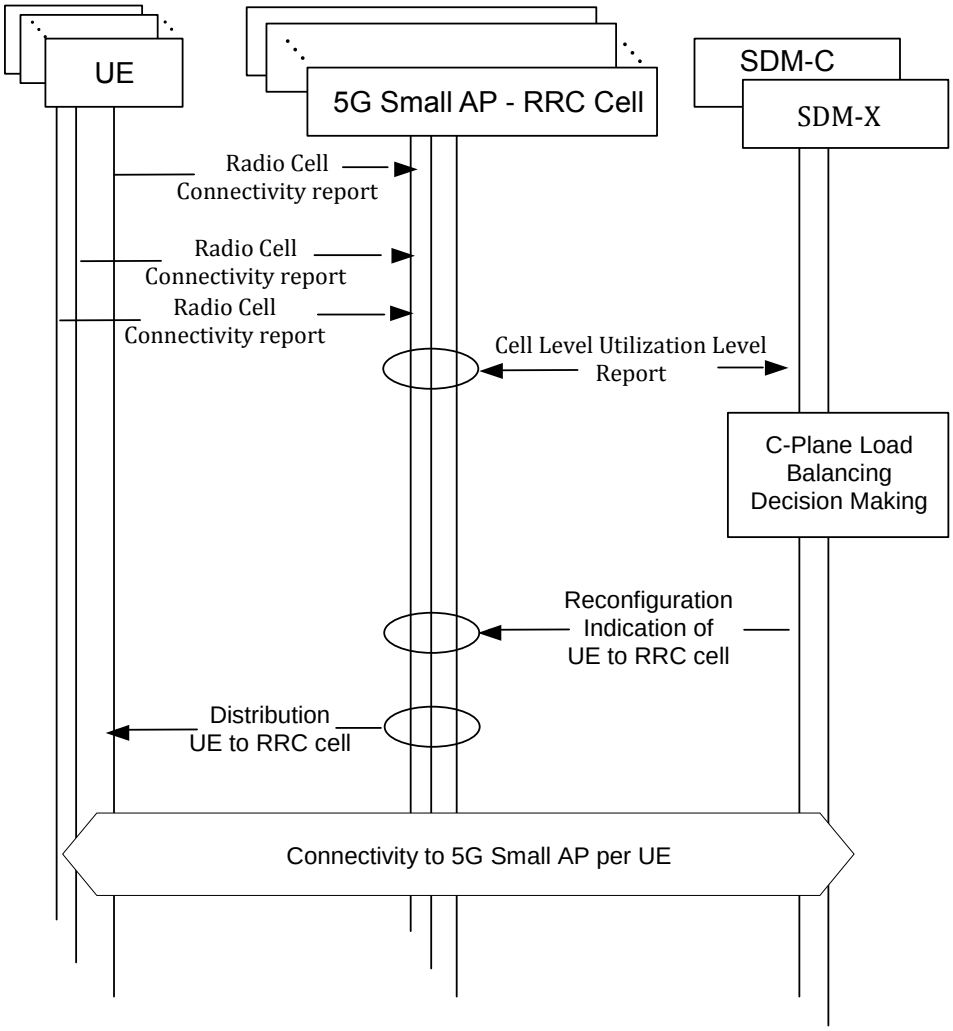


Figure 8-15: MSC for proposed SDM-C assisted load balancer

Table 8-14: Process for proposed SDM-C assisted load balancer

Actors:	**Triggering actor:** • Load information in small cells, required load per UE agent **Involved actors:** • 5G network controller SDM-C(X)/O functions
Preconditions:	• Required utilization per UE agent and connectivity in different small cells is reported to the SDM-C (c-layer)

	• UE agent is becoming active and requires a service session (c-layer)
Postconditions:	• Optimised load configuration per small cell controlled by the SDM-C via a macro-base station. Extensions can be implemented in terms of taking into account QoS per UE agent session. **We note that**: depending on the selecting frequency band it might be the case that control over the UE taken by the SDM-X controller.
Normal Course of Events:	1. UE agents connectivity per a candidate set of small cells reported to the SDM-C via a macro-base station or (depending on implementation) via connectivity to a small cell based on link gain. Aggregated information collected at the SDM-C feeds into an optimization algorithm (an optimal approach is proposed but simple heuristics can also be implemented at the SDM-C) that provide optimal allocation of UE agents per small cell with respect to load balancing across a pre-defined set of small cells (c-layer)
Special Requirements:	• This is a stateful process, so state information is required by each UE agent and therefore state must be removed either by a notification or by implementing a soft state that expires without a refresh.
Key Assumptions:	• The process can be envisioned as being adaptive and on-demand but we note that the optimization problem at hand requires that a number of requests to be handled together, i.e., in a bulk manner. This is in contrast with an on-line operation where requests are handled one by one, in an as they arrive fashion. UE agents in addition to the load scenario can be enhanced to take into account other parameters including QoS and/or QoE. • UE agents able to report to the network controller (SDM-C) their load requirements (an enrich set of parameters can also be envisioned such as those related to a specific level of QoS support for example)
Notes and Issues:	• Proposed scheme provides an optimal decision making that might not scale for large number of UE agent requests. Therefore, a heuristic algorithm might be required to provide sub optimal but competitive solutions (c-layer load reduction).

8.11. Capabilities involving geolocation information

There are numerous scenarios where geolocation information might be used to enhance 5G network capabilities, ranging from those acting solely within the mobile network (in which case there might be a strong case for the GDB being consumed into other functionality, such as the SDM-O) to those where the GDB will likely have to exist as its own distinct functionality, e.g., under regulatory-approved spectrum sharing schemes such as LSA, TVWS, etc. The intention here is not to be constrained solely to one particular usage of the GDB, such as LSA or TVWS, but to leave open the many possibilities that can be realised through such capability. The case is presented generically here for that reason. The many possibilities for the use of geolocation information and the GDB are covered in analysis/discussion in Section 7.11.

The following table and Figure 8-16present the process analysis and message sequence chart for cases involving the use geolocation information via the GDB. Although such capability can be used for many purposes, one example here is where the UE triggers a request for radio resources based on a traffic update, and this request is forwarded (and added to) by the network elements/functions. The resource decision in some cases could be made solely by the GDB (e.g., in TV White Spaces (TVWS), resource allocation); in other cases the GDB might augment information before forwarding the SDM-O for the decision (e.g., in cases where the intention is to combine/aggregate resources such as conventional licensed mobile with LSA or TVWS resources, among many other examples). There are also various monitoring, update (status report) and other aspects covered in this context and the signalling charge of Figure 8-16.

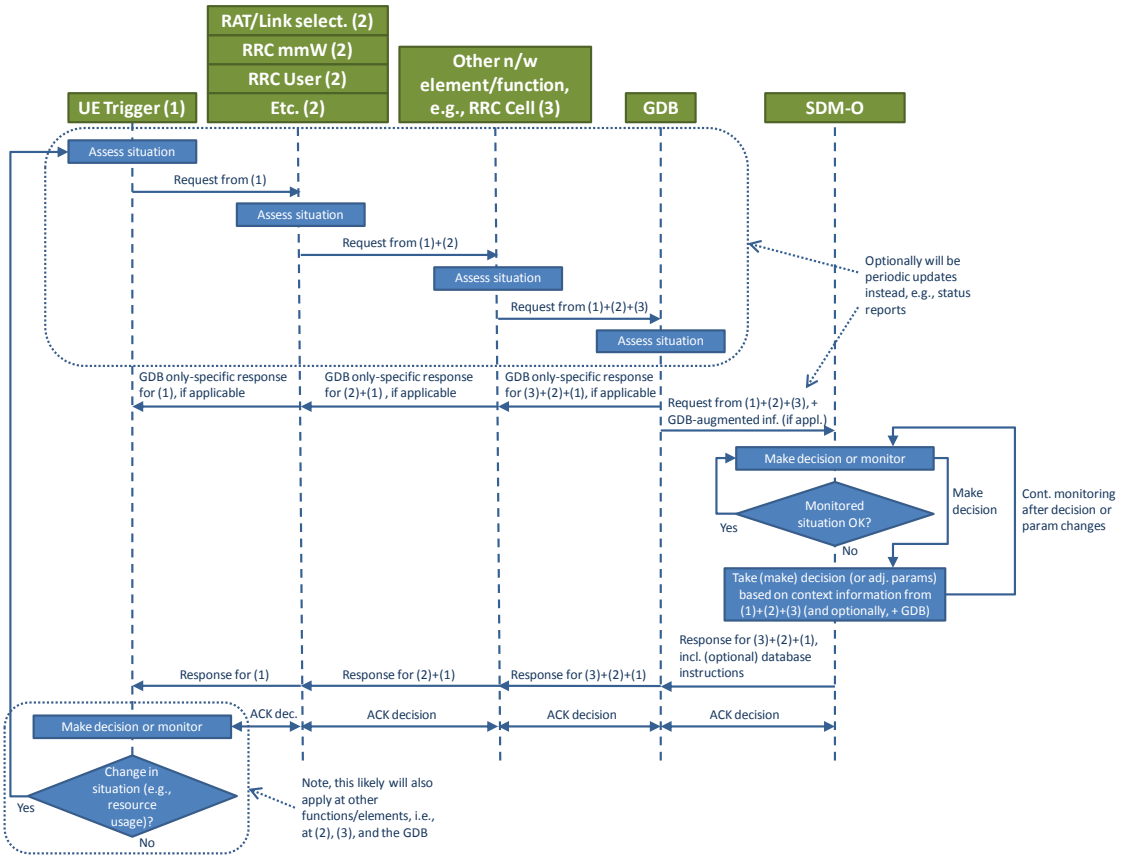


Figure 8-16: Generic MSC for capabilities involving geolocation information and the GDB

A further scenario applying to this signalling chart could be where the UE is requesting information on connection opportunities in the area, such that it can take advantage of them. The GDB can make or assisting analysis of such connection opportunities before forwarding the result to the UE or other elements/functions in the network.

Table 8-15: Process for geolocation (GDB) functionality

Actors:	**Triggering actor:** • UE (might also be an element higher-up on the network, e.g., for less fine-grain resource usage decisions are under their scope). **Involved actors:** • GDB • RRC User • RRC Cell • RRC mmW • RAT/Link selection • SON • MAC UE • SDMC • Others
Preconditions:	• UEs and other network elements/functions involved already have rudimentary Internet (or network) connectivity.
Postconditions:	• Several options (again trying to remain generic): • Resource usages are updated (perhaps across multiple levels in the network chain),

	• A better understanding of connection options is made available to UEs, • A better network resource management is achieved at higher levels (e.g., among cells and the range of bands available), • Better overall usage of available resources (spectrum efficiency) is realised.
Frequency of Use:	• Depends on the mobility/dynamics of UEs and other network elements. E.g., reassessment could be triggered if UE has moved more than 100 m from position where resources were last assessed (depends on scenario). • In many scenarios may also depend on the frequency/timescale with which the situation (e.g., resource usage) in the network is likely to change.
Normal Course of Events:	1. The UE assesses the situation (e.g., traffic requirements) and sends request based on that, in conjunction with its geolocation information 2. Other network elements/functions higher in the chain may add to this request. 3. The GDB calculates allowed resources, or even manages resources in some scenarios. 4. The GDB may forward information to the SDM-O which may play a part in managing those resources which have to be dealt with within the scope of the SDM-O (e.g., computational resources, handled based on the geolocation of their availabilities and packaged/forwarded by the GDB). 5. Resource responses are sent back to the UE and perhaps in some scenarios other network elements/functions. 6. The implementation of the decision by the UE (or in some cases other network elements) is ACKnowledged/confirmed with the GDB, and in some cases the SDM-O; it may be the case that the UE implements its own decision *within the constraints* of the response from the GDB/SDM-O, and feeds back information on its decision in the "ACK". For example, it may choose a lower transmission power than the upper allowed limit returned in the response from the GDB/SDM-O.
Alternative Courses:	1. There are numerous alternatives. One, however, could be that many of the decisions involving geolocation are made solely at the GDB, taking into account policies that have been derived at the SDM-O and conveyed to the GDB. E.g., the SDM-O's involvement is solely through policies.

9. References

[22.101] "Service aspects; Service principles." 3GPP TS 22.101, v14.0.0, June 2015.

[23.246] 3GPP TS 23.246, "Multimedia Broadcast/Multicast Service (MBMS); Architecture and functional description (Release 13)", January 2016.

[23.401] 3GPP TS 23.401 General Packet Radio Service (GPRS) enhancements for Evolved Universal Terrestrial Radio Access Network (E-UTRAN) access, v.13.4.0, 15-09-2015

[28.500] 3GPP TS 28.500, Management Concept, Architecture and Requirements for Mobile Network that include Virtualised Network Functions, Release14, Oct. 2015.

[29.305] 3GPP technical 29.305 "Interworking function between map based and diameter based interfaces (release 8)".

[31.130] 3GPP TS 31.130, "(U)SIM Application Programming Interface (API); (U)SIM API for Java™ Card", Release13, 2016.

[32.130] "Telecommunication management; Network sharing; Concepts and requirements." 3GPP TS 32.130, v12.0.0, Dec. 2014.

[32.842] 3GPP TR 32.842, Telecommunication management; Study on network management of virtualised networks, Release13, Sep. 2015.

[36.104] 3GPP TS 36.104, "Evolved Universal Terrestrial Radio Access (E-UTRA); Base Station (BS) radio transmission and reception (Release 13)", January 2016.

[36.211] 3GPP TS 36.211, "Evolved Universal Terrestrial Radio Access (E-UTRA); Physical channels and modulation (Release 13)", January 2016.

[36.212] 3GPP TS 36.212, "Evolved Universal Terrestrial Radio Access (E-UTRA); Multiplexing and channel coding (Release 13)", January 2016.

[36.213] 3GPP TS 36.213, "Evolved Universal Terrestrial Radio Access (E-UTRA); Physical layer procedures (Release 13)", January 2016.

[36.822] 3GPP 36.822 " LTE Radio Access Network (RAN) enhancements for diverse data applications", Release 11," 2012.

[5GN-D31] 5G NORMA, "Functional network architecture and security requirements", Deliverable 3.1, December 2015.

[ABA+16] D. Aziz, H. Bakker, A. Ambrosy, and Q. Liao, "Signalling minimisation framework for short data packet transmission in 5G," accepted for publication at IEEE Vehicular Technology Conference, Montreal, Canada, September 2016.

[ABK+10] A. Alexiou, C. Bouras, V. Kokkinos, A. Papazois and G. Tsichritzis, "Efficient MCS selection for MBSFN transmissions over LTE networks," Wireless Days (WD), 2010 IFIP, Venice, 2010, pp. 1-5.

[AGA+16] D. Aziz, J. Gebert, A. Ambrosy, H. Bakker and H. Halbauer, "Architecture Approaches for 5G Millimetre Wave Access Assisted by 5G Low-Band using Multi-Connectivity", submitted to Globecom Workshop 2016

[And13] J. G. Andrews, "Seven ways that HetNets are a cellular paradigm shift," *IEEE Commun. Mag.*, vol. 51, no. 3, pp. 136–144, Mar. 2013.

[BCJ+12] S. Bhaumik, S. P. Chandrabose, M. K. Jataprolu, G. Kumar, A. Muralidhar, P. Polakos, V. Srinivasan, and T. Woo, Cloudiq: A framework for processing

base stations in a data center, In Proceedings of the 18th Annual International Conference on Mobile Computing and Networking, Mobicom '12, (New York, NY, USA), pp. 125-136, ACM, 2012

[BH09] Y. Bejerano, S. Han, Cell Breathing Techniques for Load Balancing in Wireless LANs, IEEE Transactions on Mobile Computing, vol. 8, pp. 735-740, June 2009

[BHR+14] C. N. Barati, S. A. Hosseini, S. Rangan, P. Liu, T. Korakis, and S. S. Panwar, "Directional cell search for millimeter wave cellular systems," in Signal Processing Advances in Wireless Communications (SPAWC), Jun. 2014, pp. 120 - 124.

[BLR06] T. Bu, L. Li, and R. Ramjee, "Generalised Proportional Fair Scheduling in Third Generation Wireless Data Networks," in Proc. of IEEE INFOCOM, Barcelona, Spain, April 2006.

[BMK+12] M. Bansal, J. Mehlman, S. Katti, and P. Levis, Openradio:a programmable wireless dataplane, ACM HotSDN '12, pages 109-114, New York, NY, USA, 2012

[BOS+14] C.J. Bernardos, A. de la Olivia, P. Serrano, A. Banchs, L.M. Contreras, H. Jin, J.C. Zuniga, "An Architecture for Software Defined Wireless Networking", IEEE Wireless Communications Magazine, Vol.21, No.3, Jun. 2014.

[C+14] A. Checko, et al., "Cloud RAN for Mobile Networks - a Technology Overview," Communications Surveys Tutorials, IEEE, Sept 2014.

[CBR11] C. Chen, F. Baccelli, L. Roulet, Joint Optimisation of radio resources in small and macro cell networks, in Proceedings IEEE Vehicular Technology Conference, pp. 1-5, May 2011

[CCC+15] L. Cano, A. Capone, G. Carello, and M. Cesana, "Evaluating the Performance of Infrastructure Sharing in Mobile Radio Networks," in Proc. of IEEE ICC, London, UK, June 2015.

[CFY04] Y. Chen, T. Farley, and N. Ye, "QoS Requirements of Network Applications on the Internet," *Information-Knowledge-Systems Manag.*, vol. 4, no. 1, pp. 55–76, 2004.

[CIS13] Cisco Systems, "Global Mobile Data Traffic Forecast Update, 2012 - 2017," Cisco Systems, California, USA, Report, 2013.

[CPRI] Common Public Radio Interface (CPRI); Interface Specification, October 2015. http://www.cpri.info/

[CRS71] Y.S. Chow, H. Robbins and D. Siegmund, "Great Expectations: The Theory of Optimal Stopping", 1971

[CSG+13] X. Costa-Perez et al., "Radio access network virtualisation for future mobile carrier networks," IEEE Communications Magazine, vol. 51, no. 7, pp. 27-35, July 2013.

[DDM+13] Dötsch, U., Doll, M., Mayer, H.-P., Schaich, F., Segel, J. and Sehier, P. (2013), Quantitative Analysis of Split Base Station Processing and Determination of Advantageous Architectures for LTE. Bell Labs Tech. J., 18: 105-128. doi: 10.1002/bltj.21595

[DH+06] D. Niyato, E. Hossain, A Cooperative game framework for bandwidth allocation in 4G heterogeneous wireless networks, in: IEEE International Conference on Communications, 2006. ICC'06, vol. 9, June 2006, pp. 4357–4362

[DMR+15] Da Silva, I.; Mildh, G.; Rune, J.; Wallentin, P.; Vikberg, J.; Schliwa-Bertling, P.; Rui Fan, "Tight Integration of New 5G Air Interface and LTE to Fulfill 5G Requirements," in Vehicular Technology Conference (VTC Spring), 2015 IEEE 81st, vol., no., pp.1-5, 11-14 May 2015.

[DNS] Data Never Sleeps 2.0, http://www.domo.com/learn/infographic-data-never-sleeps-2

[ECC15] European Commission Communication, "Promoting the shared use of radio spectrum resources in the internal market," September 2012, accessible at https://ec.europa.eu/digital-agenda/sites/digital-agenda/files/com-ssa.pdf, accessed November 2015.

[ETSI14] ETSI, "White Space Devices (WSD); Wireless Access Systems operating in the 470 MHz to 790 MHz frequency band; Harmonised EN covering the essential requirements of article 3.2 of the R&TTE Directive," v1.1.1, April 2014.

[ETSI-MEC] ETSI, Mobile-Edge Computing - Introductory White paper, Sep. 2014.

[F+08] T. Frisanco et al., "Infrastructure sharing and shared operations for mobile network operators From a deployment and operations view," in Proc. of IEEE NOMS, Salvador, Brazil, April 2008.

[F5G-D31] FANTASTIC-5G, "Preliminary results for multi-service support in link solution adaptation", Deliverable 3.1, May 2016.

[F5G-D41] FANTASTIC-5G, "Preliminary results for multi-service support in multi-node/multi-antenna solution adaptation", Deliverable 4.1, May 2016.

[FCC10] FCC, "In the Matter of Unlicensed Operation in the TV Broadcast Bands, Additional Spectrum for Unlicensed Devices Below 900 MHz and in the 3 GHz Band, Second Memorandum, Opinion and Order," September 2010 (see also Third MO&O from April 2012).

[FCC15] FCC, "3.5 GHz Band / Citisens Broadband Radio Service: Report and Order and Second Further Notice of Proposed Rulemaking," April 2015. Accessible at https://www.fcc.gov/document/citisens-broadband-radio-service-ro, accessed July 2015.

[FRH+13] A. Ford, C. Raiciu, M. Handley, and O. Bonaventure, "Tcp extensions for multipath operation with multiple addresses," Jan. 2013, iETF RFC 6824.

[GLK14] A. Gudipati, L. E. Li, and S. Katti, "RadioVisor: A Slicing Plane for Radio Access Networks," in Proc. of HotSDN, Chicago, IL, Aug. 2014.

[GPL+13] Gudipati, A., Perry, A., Li, L. E. and Katti, S., SoftRAN: software defined radio access network, In Proc. of ACM SIGCOMM workshop HotSDN '13. ACM, New York, NY, USA, 25-30

[GSA15] A. G. Gotsis, S. Stefanatos, and A. Alexiou, “Ultra Dense Networks: The new wireless frontier for enabling 5G access,” CoRR, vol. abs/1510.05938, 2015. [Online]. Available: http://arxiv.org/abs/1510.05938

[GSM15] GSM Association, "GSMA Public Policy Position: Licensed Shared Access (LSA) and Authorised Shared Access (ASA)," February 2013. Accessible at http://www.gsma.com/spectrum/licensed-shared-access-lsa-and-authorised-shared-access-asa, accessed February 2015.

[GVP+14] A. Gember-Jacobson, R. Viswanathan, C. Prakash, R. Grandl, J. khalid, S. Das and A. Akella, "OpenNF: Enabling Innovation in Network Function Control", ACM SIGCOMM, 2014

[HD15] O. Holland, M. Dohler, "Geolocation-Based Architecture for Heterogeneous Spectrum Usage in 5G," IEEE Globecom 2015 Workshops, San Diego, CA, USA, December 2015.

[HN+14] Himayat, Nageen, et al. "Multi-radio heterogeneous networks: Architectures and performance."Computing, Networking and Communications (ICNC), 2014 International Conference on. IEEE, 2014.

[HQG+12] J. Huang, F. Qian, A. Gerber, Z. M. Mao, S. Sen, and O. Spatscheck, "A close examination of performance and power characteristics of 4G LTE networks," in Proceedings of the 10th International Conference on Mobile Systems, Applications, and Services, ser. MobiSys '12. New York, NY, USA: ACM, 2012, pp. 225–238. [Online]. Available: http://doi.acm.org/10.1145/2307636.2307658

[HWFGQ15] O. Holland, S. Wong, V. Friderikos, Y. Gao, Z. Qin, "Virtualised Sub-GHz Transmission Paired with Mobile Access for the Tactile Internet," IEEE ICT 2016, Thessaloniki, Greece, May 2016.

[HYP+12] J. Han-Shin, J. S. Young, X. Ping, J. Andrews, Heterogeneous Cellular Networks with Flexible Cell Association: A Comprehensive Downlink SINR Analysis, IEEE Transactions on Wireless Communications, vol.11, no.10, pp.3484-3495, October 2012

[IEEE1900.6] IEEE 1900.6, http://grouper.ieee.org/groups/dyspan/6, accessed November 2015.

[IJOIN-D22] iJOIN, "Definition of PHY layer approaches that are applicable to RANaaS and a holistic design of backhaul and access network", Deliverable 2.2, October 2014.

[IJOIN-D31] iJOIN, "iJOIN, "Final report on MAC/RRM state-of-the-art, Requirements, scenarios and interfaces in the iJOIN architecture" November 2013.

[IJOIN-D32] iJOIN D3.2 Definition of MAC and RRM approaches for RANaaS and a joint backhaul/access design 31-10-2014, http://www.ict-ijoin.eu/deliverables/

[IJOIN-D51] iJOIN report D5.1 "Revised definition of requirements and preliminary definition of the iJOIN architecture", October 2013, http://www.ict-ijoin.eu/

[IJOIN-D52] iJOIN D5.2 Final Definition of iJOIN Requirements and Scenarios 30-11-2014, http://www.ict-ijoin.eu/deliverables/

[ITWS15] Ofcom, "Implementing TV White Spaces," statement, February 2015. Accessible at http://stakeholders.ofcom.org.uk/consultations/white-space-coexistence/statement, accessed July 2015.

[K+12b] R. Kokku et al., "NVS: A Substrate for Virtualizing Wireless Resources in Cellular Networks," IEEE/ACM Transactions on Networking, vol. 20, no. 5, pp. 1333-1346, Oct. 2012.

[KBT+10] Khandekar, N. Bushan, J. Tingfang and V. Vanghi, "LTE-Advanced: Heterogeneous Networks", European Wireless Conference, pp. 978-982, April 2010.

[KC15] S. Khatibi and L. M. Correia, "A model for virtual radio resource management in virtual RANs," *EURASIP J. Wirel. Commun. Netw.*, vol. 2015, no. 1, p. 68, 2015.

[KMZ+13] R. Kokku, R. Mahindra, H. Zhang, S. Rangarajan, "CellSlice: Cellular Wireless Resource Slicing for Active RAN Sharing", 5th International Conference on Communication Systems and Networks (COMSNETS), Bangalore, Jan. 2013.

[LCL+14] Y. L. Lee, T. C. Chuah, J. Looh,; A. Vinel, "Recent Advances in Radio Resource Management for Heterogeneous LTE/LTE-A Networks", IEEE Communications Surveys & Tutorials, vol. 16, no. 4, pp. 2142- 2180, Nov- 2014.

[LGL+13] Y. Lin, Y. Gao, Y. Li, X. Zhang, and D. Yang, "QoS aware dynamic uplink-downlink reconfiguration algorithm in TD-LTE HetNet," *2013 IEEE Globecom Work. GC Wkshps 2013*, pp. 708–713, 2013.

[LLK+] Lars lindbom, Robert love, Sandeep Krishnamurthy, Chunhai Yao, Nobuhiko Miki, Vikram Chandrasekhar, "Enhanced Inter-cell Interference Coordination for Heterogeneous Networks in LTE-Advanced: A Survey".

[LPY08] L. Li, M. Pal, and Y. R. Yang, "Proportional fairness in multi-rate wireless LANs," in Proc. of IEEE INFOCOM, Phoenix, AZ, April 2008.

[LSW+11] M. Luby, A. Shokrollahi, M. Watson, T. Stockhammer, and L. Minder, "Raptorq forward error correction scheme for object delivery," Aug. 2011, iETF RFC 6630.

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

[MCWSD15] Ofcom, "Manually Configurable White Space Devices," Consultation, February 2015, accessible at http://stakeholders.ofcom.org.uk/consultations/manually-configurable-wsds, accessed November 2015.

[MDU] Make Data Useful, http://www.gduchamp.com/media/StanfordDataMining.2006-11-28.pdf

[MET-D11] METIS2020 D1.1 Scenarios, requirements and KPIs for 5G mobile and wireless system 30-04-2013, https://www.metis2020.com/documents/deliverables/

[MET-D64] METIS 2020 D6.4 Final report on architecture 31-01-2015, https://www.metis2020.com/documents/deliverables/

[METII-R21] ICT-671680-METIS-II / R2.1: RAN Design Guidelines, Sep 2015

[MKH+13] Mahindra, R., Khojastepour, M.A., Honghai Zhang, and Rangarajan, S., Radio Access Network sharing in cellular networks, IEEE Network Protocols (ICNP) 2013, pp.1,10, 7-10 Oct. 2013

[MVA14] I. Malanchini, S. Valentin, and O. Aydin, "Generalised resource sharing for multiple operators in cellular wireless networks," in Proc. of IWCMC, Nicosia, Cyprus, Aug. 2014.

[MKZ+13] R. Mahindra, M. Khojastepour, H. Zhang, and S. Rangarajan, "Radio Access Network sharing in cellular networks," in Proc. of IEEE ICNP, Goettingen, Germany, Oct. 2013.

[NFV-002] ETSI GS NFV-002 Architectural Framework, v1.2.1, Dec. 2014.

[NGMN15] NGMN Alliance, Next Generation Mobile Networks 5G White Paper V1.0, NGMN 5G Initiative, February 2015

[NOK-WPa] LTE-Advanced Evolution in Releases 12 - 14. 'New services to pave the way to 5G' Nokia Networks.

[NOK-WPb] Ten key rules for 5G deployment: Enabling 1 Tbit/s/km2 in 2030, Nokia Networks White Paper, April 2015.

[NOKWP] LTE networks for public safety services, Nokia Networks white paper.

[OY02] L. Ong and J. Yoakum, "An introduction to the stream control transmission protocol," May 2002, iETF RFC 3286.

[PBF+16] K. I. Pedersen, G. Berardinelli, F. Frederiksen, P. Mogensen and A. Szufarska, "A flexible 5G frame structure design for frequency-division duplex cases," in IEEE Communications Magazine, vol. 54, no. 3, pp. 53-59, March 2016. doi:2010.1109/MCOM.2016.7432148

[PCAST12] President's Council of Advisors on Science and Technology (PCAST), "Report to the President: Realizing the Full Potential of Government-Held Spectrum to Spur Economic Growth," July 2012. Accessible at http://www.white-house.gov/sites/default/files/microsites/ostp/pcast_spectrum_report_fi-nal_july_20_2012.pdf, accessed February 2015.

[PKB+11] K. Piamrat, A. Ksentini, J. M. Bonnin and C. Viho, "Radio resource manage-ment in emerging heterogeneous wireless networks", Computer Communica-tions, vol. 34, no. 9, pp. 1066-1076, June 2011.

[PLS15] V. Pauli, Y. Li and E. Seidel, "Dynamic TDD for LTE-A", NOMOR Research GmbH white paper, Sep. 2015

[PMD+15] E. Pateromichelakis, A. Maeder, A. De Domenic, R. Fritzsche, P. de Kerret and J. Bartelt, "Joint RAN/Backhaul Optimisation in Centralised 5G RAN". Networks and Communications (EuCNC), 2015 European Conference on. June 29 2015-July.

[R+16] A. Ravanshid et al., "Multi-connectivity functional architectures in 5G," 2016 IEEE International Con-ference on Communications Workshops (ICC), Kuala Lumpur, 2016, pp. 187-192.

[RS-12] LTE-Advanced (3GPP Release12) Technology Introduction, Rohde & Schwarz LTE-Advanced (3GPP Release12) Technology Introduction. https://cdn.rohde-schwarz.com/pws/dl_downloads/dl_application/applica-tion_notes/1ma252/1MA252_WP_LTE_Rel12_2E.pdf

[RWS-12] 3GPP RWS-120010 WS Docomo, Requirement, candidate Solutions and Technology Roadmap for LTE Release-12 onward, 3GPP Workshop on Re-lease 12 and on-wards Ljubljana, Slovenia, June 11-12, 2012

[SCF-15a] Small Cell Forum, Small Cell Virtualisation Functional Splits and Use Cases, Document 159.05.1.01, June 2015.

[SCF-15b] Small Cell Forum, Network Aspects of Virtualised Small Cells, Release 5.1, Document 161.05.1.01, Jun. 2015.

[SEK+12] Z. Shen, A. Khoryaev, E. Eriksson, and X. Pan, "Dynamic uplink-downlink configuration and interference management in TD-LTE," *IEEE Commun. Mag.*, vol. 50, no. 11, pp. 51–59, Nov. 2012.

[SDM+12] SaiShankar N, Debashis Dash, Hassan El Madi, and Guru Gopalakrishnan Tensorcom, "WiGig and IEEE 802.11ad For Multi-Gigabyte-Per-Second WPAN and WLAN", available at http://arxiv.org/ftp/arxiv/pa-pers/1211/1211.7356.pdf.

[Sig_Storm] http://www.diametriq.com/wp-content/uploads/downloads/2013/01/A-Storm-is-Brewing.pdf

[SMD15] L. Sanguinetti, A. L. Moustakas and M. Debbah, "Interference Management in 5G Reverse TDD HetNets with Wireless Backhaul: A Large System Analy-sis", IEEE Journals on Selected Areas in Communications (JSAC), vol. 33, no. 6, pp. 1187-1200, June 2015

[SMR+15] Icaro Da Silva, Gunnar Mildh, Johan Rune, Pontus Wallentin, Jari Vikberg, Paul Schliwa-Bertling, Rui Fan, "Tight Integration of New 5G Air Interface

and LTE to Fulfill 5G Requirements," 2015 IEEE 81st Vehicular Technology Conference (VTC Spring), Glasgow, May 2015. doi:10.1109/VTCSpring.2015.7146134

[SMS+16] Icaro Leonardo Da Silva, Gunnar Mildh, Mikko Säily, Sofonias Hailu, "A novel state model for 5G radio access networks", Workshop on 5G RAN Design (5G RAN), ICC16, May 2016.

[SPS+16] K. Samdanis, X. Perez-Costa, V. Sciancalepore, "From Network Sharing to Multi-tenancy: The 5G Network Slice Broker". IEEE Communication Magazine - Communication Standards, February 2016.

[SPW+96] Katia Sycara, Anandeep Pannu, Mike Williamson, and Dajun Seng, Keith Detker, "Distributed Intelligent Agent", IEEE Expert, Volume: 11, Issue: 6, December 1996, Page(s): 36 - 46.

[SSP+16] K. Samdanis, R. Shirvastava, A Prasad, D. Grace, and X. Costa-Perez, Samdanis, TD-LTE virtual cells: An SDN architecture for user-centric multi-eNodeB elastic resource management. *Computer Communications*, *83*, 1-15.

[SW14] Frank Schaich, Thorsten Wild; "Relaxed synchronisation support of universal filtered multi-carrier including autonomous timing advance", Wireless Communications Systems (ISWCS), 2014 11th International Symposium on. IEEE, 2014. S. 203-208.

[SWA16] F. Schaich, T. Wild, R. Ahmed, "Subcarrier spacing - how to make use of this degree of freedom," IEEE Vehicular Technology Conference Spring, Nanjing, China, May 2016

[SWC14] F. Schaich, T. Wild, and Y. Chen, "Waveform contenders for 5G - suitability for short packet and low latency transmissions," in Vehicular Technology Conference (VTC Spring), 2014 IEEE 79th, May 2014, pp.1–5.

[SWS15] S. Saur, A. Weber, and G. Schreiber, "Radio access protocols and preamble design for machine-type communications in 5G," in Signals, Systems and Computers, 2015 49th Asilomar Conference on, Nov 2015, pp. 3-7.

[TAT] The ABCs of TCP/IP, http://www.ciscopress.com/articles/article.asp?p=377101

[Tech-Lib] http://developer.att.com/technical-library/network-technologies/long-term-evolution

[TR+15] Trivisonno, R., et al. "SDN?based 5G mobile networks: architecture, functions, procedures and backward compatibility." Transactions on Emerging Telecommunications Technologies 26.1 (2015): 82-92.

[TWSD15] Ofcom TV White Space Databases, https://tvws-databases.ofcom.org.uk, accessed November 2015.

[WH+15] Wang, Hucheng, et al. "SoftNet: A software defined decentralised mobile network architecture toward 5G Network", IEEE 29.2 (2015): 16-22.

[WHB15] N. Wang, E. Hossain, and V. K. Bhargava, "Backhauling 5G small cells: A radio resource management perspective," *IEEE Wirel. Commun.*, vol. 22, no. 5, pp. 41–49, Oct. 2015.

[WLM+05] A. Wilson, A. Lenaghan, R. Malyan, Optimising wireless access network selection to maintain QoS in heterogeneous wireless environments, Wireless Personal Multimedia Communications 2005. WPMC'05, 18–22 September 2005.

[WWRF] http://www.wwrf.ch/files/wwrf/content/files/publications/outlook/Outlook9.pdf

[XBC+05] X. Yang, J. Bigham, L. Cuthbert, Resource management for service providers in heterogeneous wireless networks, in: IEEE Wireless Communications and Networking Conference 2005, vol. 3, 13–17 March 2005, pp. 1305–1310

[YVU+14] Yazıcı, Volkan, Ulas C. Kozat, and M. Oguz Sunay. "A new control plane for 5G network architecture with a case study on unified handoff, mobility, and routing management." Communications Magazine, IEEE 52.11 (2014): 76-85.

[Y+13] Q. Ye et al., "User Association for Load Balancing in Heterogeneous Cellular Networks," IEEE Transactions on Wireless Communications, vol. 12, no. 6, pp. 2706-2716, June 2013.

[ZZG+13] S. Zhang, Z. Zhao, H. Guan, D. Miao, and H. Yang, “Statistics of RRC state transition caused by the background traffic in LTE networks,” in Wireless Communications and Networking Conference (WCNC), 2013 IEEE, April 2013, pp. 912–916.