Broadcast and Multicast Communication Enablers for the Fifth-Generation of Wireless Systems

Deliverable D3.3

RAN Logical Architecture and Interfaces for 5G-Xcast

Version v2.0 Date: 2019/02/28

Document properties:

Grant Number:	761498
Document Number:	D3.3
Document Title:	RAN Logical Architecture and Interfaces for 5G-Xcast
Editor(s):	Mikko Säily (NOK), Carlos Barjau (UPV)
Authors:	BBC (David Vargas), IRT (Jordi J. Gimenez), NOK (Mikko Säily, David Navrátil, Athul Prasad), NOM (Fasil B. Tesema), O2M (Peter Sanders), SEUK (Wei Guo), UNIS (De Mi), UPV (Carlos Barjau)
Reviewers:	Christian Mannweiler
Contractual Date of Delivery:	2019/02/28
Dissemination level:	PU[1]
Status:	Final
Version:	2.0
File Name:	D3.3_v2.0

Abstract

Deliverable D3.3 provides analysis and technical solutions developed in the 5G-Xcast project considering the use cases and technical requirements discussed in deliverables D2.1 and D2.2. This document designs the 5G RAN logical architecture to deploy flexible multicast radio access network that can support the same services as existing multicast/broadcast networks as well as removing limitations concerning the specifications in LTE Rel-14. This document further evolves the support for high data rate media content, with a seamless switching between unicast and multicast, and develops solutions to deliver terrestrial broadcast services reusing the designed dynamic multicast RAN architecture. It also provides concepts and procedures to deploy multicast in dynamic geographical areas.

Keywords

3GPP Radio Access Network Architecture, LTE-Advanced Pro, Point-to-Multipoint, Point-to-Point, Radio Access Network, 5G-Xcast RAN Technical Requirements, Use Cases, Functional Split, Interfaces, RAN Multicast, Protocol, Content Synchronization, Media & Entertainment vertical, Public Warning vertical, Automotive vertical, IoT vertical.

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

PU = Public

Revision History

Revision	Date	Issued by	Description
1.0	2018/08/31	Mikko, Carlos	Interim Report
2.0	2019/02/28	Mikko, Carlos	Final report

Disclaimer

Executive Summary

The first 3GPP release of 5G technology (Rel'15), also known as NR (New Radio), was completed in 2018. This release of standards is focused on unicast communications in the core network and point-to-point (PTP) transmissions in the Radio Access Network (RAN). While the unicast communication is the traditional way to deliver content in cellular networks, the radio resources needed to deliver a common IP multicast content to many users will grow linearly with the number of users, causing high load in the air interface. The support of multicast/broadcast is needed for point-to-multipoint (PTM) transmissions in vertical sectors, such as Automotive, Airborne Communications, Internet-of-Things, Media & Entertainment, and Public Warning & Safety systems. In 5G multicast architecture, the future trends in technology and service design should be considered in the design process.

3GPP Rel-15 specifications have standardized the design principles of cloud-native system approach. The key architectural element in RAN design was to extend the distributed base station architecture towards flexible cloud-RAN based protocol functionality where the computing hardware pools are used to take care of the higher layer processing of user plane data traffic and control plane signalling. The split between central units (CUs) and distributed units (DUs) in 5G architecture enables dynamic adaptation of QoS functions depending on the real-time radio conditions, user density and dynamically controlled geographical area for multicast. These architecture enhancements provide a significant opportunity to design an innovative RAN architecture for multicast content in 5G-Xcast project.

The 5G multicast services are required to be supported by multitude of network deployments with increased flexibility, and dynamically adapt in areas where the demand for multicast will change significantly over the time, e.g. moving users and/or stadiums. 5G-Xcast RAN architecture will be aware of the user's interest to receive IP multicast data and therefore the developed dynamic RAN Multicast Areas (RMA) allow delivery of multicast services wherever needed without eMBMS-type of static deployment on top of the existing RAN. RMA mechanism can consider the user activity, user mobility, number of devices and their geographical distribution.

The multicast/broadcast services over multiple network cells require content synchronization to share the same content among multiple users. 5G-Xcast RAN architecture supports a multi-cell transmission using NG-RAN based synchronization method, where synchronized DUs participate to multi-cell transmission using a single CU as a point for transmission coordination. This approach enables over the air transmission of synchronized multicast/broadcast traffic while fulfilling the QoS targets defined for the traffic flows.

In this deliverable, we present novel technical developments enhancing the 5G unicast architecture to enable efficient selection between unicast and multicast/broadcast transmission modes in dynamic geographical areas using the enhanced NG-RAN architecture based on 3GPP Rel-15 and the architectural and functional enhancements developed during the 5G-Xcast project. Solutions are presented to support also the delivery of Terrestrial Broadcast services within the same 5G-Xcast architecture.

The result is a RAN architecture approach which can adapt to a wide variety of requirements and QoS targets defined according to user demand and traffic flows and network operators and service providers.

Table of Contents

List of Figures

List of Tables

List of Acronyms and Abbreviations

3GPP	3rd Generation Partnership Project
5G	5th Generation
5GC	5G Core Network
AMF	Access and Mobility Management Function
AN	Access Network
CN	Core Network
CQI	Channel Quality Indicator
CoMP	Coordinated MultiPoint
CP	Control Plane
CQI	Channel Quality Indication
CSC	Communication Service Consumer
CSI	Channel State Information
CSP	Communication Service Provider
CU	Centralized Unit
CU-C	Centralized Unit Control Plane
CU-CP	Centralized Unit Control Plane
CU-U	Centralized Unit User Plane
CU-UP	Centralized Unit User Plane
CU-MC	Centralized Unit MultiCast
DL	Downlink
DRB	Data Radio Bearer
DTT	Digital Terrestrial Television
DU	Distributed Unit
eMBB	enhanced Mobile Broadband
eMBMS	evolved Multimedia Broadcast Multicast Service
eNB	eNodeB
E-UTRAN	Evolved Terrestrial Radio Access Network
EPC	Enhanced Packet Core
EPG	Electronic Programme Guide
E1AP	E1 Application Protocol
F1AP	F1 Application Protocol
FEC	Forward Error Correction
gNB	NR NodeB
GNSS	Global Navigation Satellite System
GPRS	General Packet Radio Service
GTP-U	GPRS Tunnelling Protocol – User Plane
IGMP	Internet Group Message Protocol
IAB	Integrated Access and Backhaul
IoT	Internet Of Things
IP	Internet Protocol
ITU	International Telecommunication Union
LCM	Lifecycle Management
M&E	Media and Entertainment
MAC	Medium Access Control
MBMS	Multimedia Broadcast Multicast Service
MBSFN	MBMS over Single Frequency Networks
MCCH	Multicast Control Channel
MCE	Multi-cell/multicast Coordination Entity
MCF	Multicast Function
MC-PTM	Multi-Cell – Point-to-Multipoint
MCS	Modulation and Coding Scheme
M&E	Media and Entertainment
MIMO	Multiple Input Multiple Output
ML	Maximum Likelihood

MLD	Multicast Listener Discovery
MNO	Mobile Network Operator
MTCH	Multicast Traffic Channel
NB-IoT	Narrow-Band IoT
NF	Network Function
NFV	Network function virtualisation
NG	NextGen
NG-RAN	NextGen RAN
NGAP	NG Application Protocol
NOP	Network Operator
NR	New Radio
NSA	Non-Standalone
NSaaS	Network Slice as a Service
NSS	Network Slice Subnet
NSSI	Network Slice Subnet Instance
PBCH	Physical Broadcast Channel
PDCCH	Physical Downlink Control Channel
PDCP	Packet Data Convergence Protocol
PDSCH	Physical Downlink Shared Channel
PDU	Protocol Data Unit
PHICH	Physical HARQ Indicator Channel
PHY	Physical Layer
PNF	Physical Network Function
PRACH	Physical Random Access Channel
PRB	Physical Resource Block
PSS	Primary Synchronization Signal
PTM	Point-to-Multipoint
PTP	Point-to-Point
PUCCH	Physical Uplink Control Channel
PUSCH	Physical Uplink Shared Channel
PW	Public Warning
PWS	Public Warning System
QAM	Quadrature Amplitude Modulation
QoS	Quality of Service
RAN	Radio Access Network
RAT	Radio Access Technology
RB	Radio Bearer
Rel	Release
RF	Radio Frequency
RLC	Radio Link Control protocol
RLC-AM	Radio Link Control-Acknowledged Mode
RLC-UM	Radio Link Control-Unacknowledged Mode
RMA	RAN Multicast Area
ROM	Receive Only Mode
RRC	Radio Resource Control
RRH	Remote Radio Head
RRM	Radio Resource Management
RU	Remote or Radio Unit
S-NSSAI	Single Network Slice Selection Assistance Information
SA	Stand-Alone
SC-PTM	Single Cell – Point-to-Multipoint
SCTP	Stream Control Transmission Protocol
SD	Slice Differentiator
SDAP	Service Data Adaptation Protocol
SLA	Service Level Agreement
SMF	Session Management Function
SON	Self-Organizing Network
SST	Slice/Service Type

SYNC	Synchronization protocol
TB	Terrestrial Broadcast
TB-SA	Terrestrial Broadcast Service Area
TNL	Transport Network Layer
TR	Technical Report
TS	Technical Specification
TV	Television
UE	User Equipment
UL	Uplink
UDP	User Datagram Protocol
UP	User Plane
UPF	User Plane Function
URLLC	Ultra Reliable and Low Latency Communications
V2X	Vehicular to everything
VNF	Virtual Network Function
WP	Work Package
XnAP	Xn Application Protocol
XCF	5G-Xcast Control Plane Network Function
XRB	Xcast (Multicast) data Radio Bearer
XUF	5G-Xcast User Plane Network Function

1 Introduction

1.1 Background and scope

The 5G-Xcast project is focused on content distribution & delivery over 5G networks to address the continued growth of content services and requirements for point-to-multipoint delivery. 5G-Xcast will define a unified logical Access Network (AN) or Radio Access Network (RAN) architecture which can dynamically exploit unicast, multicast and broadcast delivery modes and be configured to distribute Terrestrial Broadcast services. 5G-Xcast technologies will be fundamental in the progression towards a converged 5G architecture and efficient in IP multicast traffic delivery over radio technologies.

5G-Xcast has defined use cases that demonstrate and specify the applicability and requirements. At the same time, 3GPP on the mobile access side has been working the first 5G specification, e.g. New Radio (NR). This document provides full evaluation of the RAN logical architecture framework for 5G-Xcast in the scope of project requirements and identified use cases.

1.2 3GPP architecture evolution to Release 15

In 3GPP standardization forum the Rel-15 specifications are being finalized focussing on the design principles of forward compatibility, control-user plane separation, lean and cloud-native system design. One of the key elements of the RAN architecture design is to support centralized processing in Cloud-RAN with protocol functionality split of NR base station, namely gNB, between central units (CUs) and distributed units (DUs). Such novel architecture enhancements provide a significant opportunity to design an innovative RAN architecture for delivering multicast content in 5G.

According to Figure 1-1, the gNB functions are split into CU and DU, where CU covers higher layer protocol functions of SDAP and PDCP, and DU covers lower layer protocol functions of RLC, MAC and PHY. In a typical Cloud-RAN deployment, the CUs are placed in a computing hardware pool and thus form the cloud. The gNBs are inter-connected through a Xn interface.

The F1 interface provides control (F1-C) and user (F1-U) plane connectivity between the CU and DU, enabling deployments with C/U-plane separation on the CU level. The E1 interface provides connectivity between the user plane CU-UP and control plane CU-CP, also enabling deployments with C/U-plane separation. The interface also provides separation between the radio network and transport network layers, while enabling the exchange of UE and non-UE associated information. When F1 is separated into F1-C on CP and F1-U on UP, consequently the Xn inter-connecting the gNBs is separated into Xn-C on CP and Xn-U on UP. A gNB-CU is further separated logically into gNB-CU-CP on CP and gNB-CU-UP on UP, e.g. with the E1AP interface and application protocol.

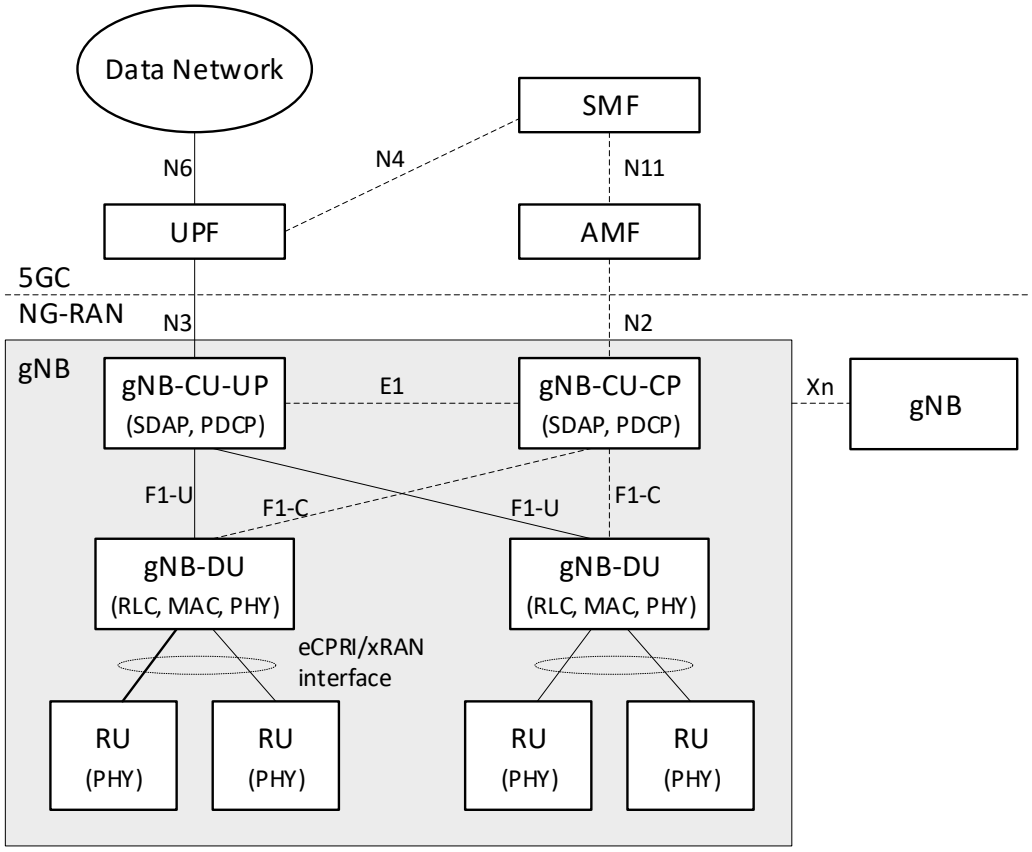


Figure 1-1: NG-RAN architecture with a CU-DU split deployment

1.3 Objectives of the RAN Logical Architecture design in 5G-Xcast

5G-Xcast WP3 is studying the RAN aspects such as the performance of the existing state-of-the-art systems, air interface, the RAN architecture and the RAT protocols. WP3 aims at designing flexible baseline RAN architecture solution based on 3GPP NG-RAN Release 15 that can fulfil the requirements and use cases developed in 5G-Xcast WP2.

1.4 Structure of the document

This document is organized as follows. The document introduces and describes the architecture evolution of MBMS up to 3GPP Release 14 including limitations of existing systems and followed by architectural requirements of the 5G-Xcast system in chapter 2. Chapter 3 defines the 5G-Xcast logical architecture including functional entities, function splits and multicast content synchronization. Chapter 4 describes the interfaces between 5G-Xcast RAN and 5GC, RAN Internal Interfaces and the interface between RAN and UE. Further Chapter 5 defines the protocol stack architecture and the procedures supporting the flexible 5G-Xcast deployment. In Chapter 6 the RAN deployment and network functions are defined, followed by RAN Slicing in Chapter 7 and the Summary and Conclusions of developed 5G-Xcast RAN Logical Architecture in Chapter 8.

2 5G-Xcast RAN Architecture Requirements

2.1 RAN Architectural Requirements from Use Cases

5G-Xcast project WP2 “Use Cases, Requirements & KPIs” specified the requirements to be used when the new techniques are being developed within 5G-Xcast. The use cases focus on applications that have the most to gain from the concepts being developed within the project, and the 5G-Xcast WP3 focuses specifically to solve the challenges of RAN logical architecture requirements. The addressed concepts which have the main impact to the technical developments in WP3 are flexible point-to-multipoint (PTM) capability, dynamic and adaptable network architecture, dynamic and seamless switching between unicast, multicast and broadcast modes or their simultaneous use.

Several requirements for the Media & Entertainment (M&E), Public Warning System (PWS) and Internet Of Things (IOT) verticals have been identified in [2]. Regarding M&E, only two requirements have several implications on the overall design of the 5G-Xcast RAN Architecture. The first one is M&E1_R7, which requires the dynamic optimization of resource allocation. To enable dynamic optimization, it is necessary to have a common delivery mode framework where PTP, Mixed-Mode and Terrestrial Broadcast can be switched on-the-fly, based on user demand and geographical location.

Considering M&E1_R20, the second requirement, it requires that the same security procedures applied to PTP transmissions should be applied in PTM ones i.e. content integrity, confidentiality, availability and non-repudiation. In order to fulfil this, a reuse (or extension if necessary) of NG-RAN protocol stack at all related interfaces in 5G-Xcast RAN is proposed. Note that, for the Terrestrial Broadcast mode, if confidentiality is enforced, the relevant encryption keys should be made available to the users via other channels e.g. Unicast for non-ROM devices.

The paradigm is different for PWS verticals, where unpredictability, both in temporal and spatial domains, define the PWS situations. In case of an alert, the critical content must be broadcasted only in the affected area, so cell granularity is needed (PWS_R4).

Cell granularity is not the only PWS requirement that affects the RAN: Keeping the same battery usage as unicast while receiving broadcast data is another requisite (PWS_R5). To accomplish this, in 5G-Xcast RAN, the use of RRC_INACTIVE state for the UEs, with low battery consumption, will be actively employed in Sections 0 and 5

Lastly, for IoT verticals, the requirements affecting the RAN are two. On the one hand, uplink reporting available for confirmation of file delivery e.g. IoT devices confirming successful firmware update. This requires that the logical multicast/broadcast channels used for 5G-Xcast RAN allow for simple feedback (ACK on file delivery, CSI). This feature will be optional when launching a Xcast Broadcast transmission. On the other hand, IoT devices which are battery-constraint are supposed to have very low active mode time in RRC_CONNECTED state, so it is necessary to notify them when to wake up, so they can receive Software Updates or other critical signalling (IoT_R2). To satisfy this requirement, a Service Announcement transport channel is needed, not only for IoT devices but also useful for Receive-Only-Mode devices, while keeping the content transparent to 3GPP part of the network, e.g. distribution of the Electronic Programme Guide (EPG).

To summarize all the needed features from the Use Cases requirements according to 5G-Xcast deliverables D2.1 [2] and D2.2 [3]:

- Common delivery mode framework to enable dynamic switching between modes
- Reuse of Unicast logic, transport and physical procedures to guarantee security constraints
- Cell granularity when deploying the Multicast/Broadcast transmission
- Maintain or reduce the battery usage compared to Point-to-Point
- Feedback channel permitted
- Service announcement channel

Note that in detailed design and practical deployment in various use cases there are more requirements and limitations to be addressed than the ones showed in this section.

2.2 Requirements from NR architecture and access technology perspective

3GPP Release 14 Study on scenarios and requirements for next generation systems [45] evaluated scenarios and requirements for (so called) Supplementary Services, which cover the main aspects for 5G-Xcast from the radio access technology perspective. Some of the requirements have major impact on the RAN architecture design, in terms of communication protocols and supporting different deployments and use cases.

Requirements impacting RAN architecture are related to support of dynamic deployment of 5G-Xcast RAN where the new RAT shall support dynamic adjustment of the Multicast/Broadcast area based on e.g. the user distribution or service requirements. The new RAT should also cover large geographical areas up to the size of an entire country in SFN mode with network synchronization and shall allow cell radii of up to 100 km if required to facilitate that objective [45]. It shall also support local, regional and national broadcast areas. The requirement of dynamic geographical areas and large nationwide areas also demands for support of concurrent delivery of both unicast and Multicast/Broadcast services to the users, as well as support for efficient multiplexing with unicast transmissions. The new RAT shall support Multicast/Broadcast services for fixed, portable and mobile UEs. Mobility up to 250 km/h shall be supported.

2.3 Architectural Limitations of LTE Release 14

2.3.1 Limitations for Unicast/Multicast/Broadcast operation

3GPP Rel-14 eMBMS RAN logical architecture covers solutions where services can be provided either using a single frequency network mode (MBSFN) or MBMS transmission using Single-Cell Point-to-Multipoint transmission [2]. In this case the MBSFN can be understood as a transmission method on a frequency layer shared with non-MBMS services or on a frequency layer dedicated for MBMS. The Rel-14 specification was evaluated in 5G-Xcast deliverable D3.1 including technical details impacting the RAN logical architecture such as RAN synchronization, RAN coverage area adaptation & extension, adaptation between unicast/broadcast multiplexing and adaptation between MBSFN and SC-PTM transmission modes.

Current MBSFN and SC-PTM solutions are based on a static configuration to deliver the broadcast/multicast traffic to predefined areas. While in 3GPP Rel-14 the MBSFN

area is statically configured regardless of user distribution, an alternative to this is to enable dynamic geographical multicast areas where the RAN can decide the best means of transmitting the multicast traffic to UEs indicating the interest of receiving multicast content. In 3GPP Rel-14 it is not possible to dynamically create SFN areas on a multi-cell level and it is not possible to adaptively use PTP, single cell PTM and multiple cell SFN/PTM in different areas of the network to deliver the same service.

When looking at the multicast content delivery from the overall system architecture perspective, the eMBMS RAN architecture consists of E-UTRAN nodes, namely the eNodeB and the Multi-cell/multicast Coordination Entity (MCE). In 3GPP Rel-14 MBMS the MCE selects the MCS so that the service level agreement of the broadcast/multicast service is met in a given coverage area. Thereafter, the Rel-14 solution is lacking radio channel feedback mechanisms to optimize radio resource allocation, optimum MCS and optimal selection of multicast or unicast data bearers. Related MCE functionality in 5G RAN architecture will be investigated to enable optimisation of use cases of broadcast/multicast and unicast and adaptation between these transmission modes.

Rel-14 eMBMS control plane signalling and user plane data packets are distributed from the EPC to E-UTRAN through dedicated interfaces M2/M3 and M1 interface respectively. The dedicated architecture of eMBMS presents additional complexity and overhead to the network, which according to mixed mode multicast requirements should be maximizing the usage of unicast RAN architecture leading to a common and flexible architecture and deployment.

The Rel-14 MBSFN architecture supports different mobility classes defined in IMT-2020, but on the other hand don't allow for dynamic Multicast area based on e.g. the user geographical distribution. That is, 5G-Xcast RAN architecture should define dynamic multicast areas where the multicast content is available regardless of the user unicast activity. The 5G-Xcast project will design dynamic RAN multicast areas which can allow selection of the most efficient transmission method, thus improving system capacity and enabling a low probability of failure to deliver the multicast data the users. Introduction of RRC_INACTIVE state in 3GPP NR Rel-15 allows definition of dynamic RAN multicast procedures, which increase the flexibility of RAN procedures and allow dynamic multicast service areas according to number of users, their geographical distribution and service requirements.

Finally, the lack of feedback mechanisms prevents Rel-14 eMBMS meeting requirements where user service usage and related feedback reports may be of interest for network optimization and/or to the service provider.

3GPP NR Release 15 has specified the new RAN logical architecture and RAN-Core interfaces. 5G-Xcast project will evaluate the Release 15 NR RAN architecture and RAN-Core interfaces, take the new scenarios and requirements into account and address the above listed limitations of RAN logical architecture design.

2.3.2 Limitations for Terrestrial Broadcast operation

3GPP Rel-14 eMBMS already provides mechanisms suitable for the configuration of different broadcast coverage areas. However, the specified framework is static in terms of service allocation and the mechanisms are mostly oriented for SFN deployments, whereas state-of-the-art Terrestrial Broadcast networks should rely on a variety of deployment options.

From Rel-14 operation modes, MBSFN is suitable for SFN deployments where SC-PTM may be an option for single-cell transmissions (or, in extension, multi-frequency networks, MFN), both enabling the allocation of Terrestrial Broadcast services up to 100% of the carrier resources. Enhanced flexibility would allow the creation and adaption of coverage areas according to operator demands or the distribution of services to several areas within e.g. a country to provide nationwide coverage.

The use of a common architecture, based on the existing unicast architecture, with the ability to just select the functions suitable for Terrestrial Broadcast distribution, will minimize the implementation complexity of the RAN.

3 5G-Xcast RAN Logical Architecture

3.1 Functional entities to support Point-to-Multipoint in 5G architecture

3.1.1 Overview of multicast traffic delivery

The high-level architecture for enabling multicast in a 5G network is shown on Figure 3-1. The User Plane Function (UPF) receives multicast data over the N6 reference point. The multicast data is transported to the RAN nodes (only one node is shown for simplicity) over core network data tunnels at reference points N3 and N9 (N9 not shown for simplicity). The high-level procedure for multicast traffic delivery consists of the access network allocating the resources based on UE's interest in receiving multicast, e.g. transmission of IGMP/MLD (Internet Group Message Protocol/Multicast Listener Discovery), PDU session modification for each UE, the RAN awareness of the UE's interested in receiving multicast data and the core network tunnel which is carrying the data.

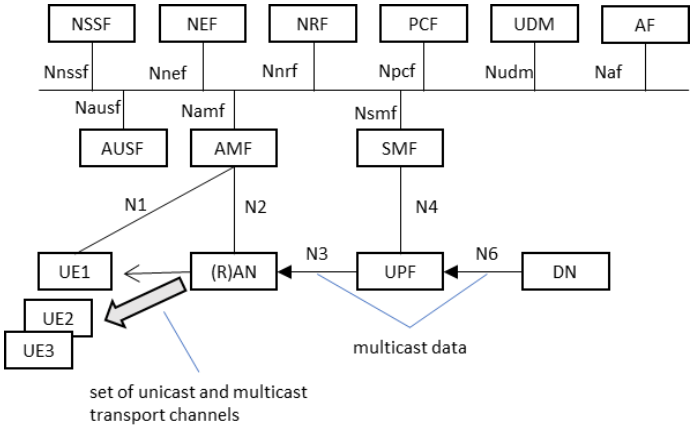


Figure 3-1: 5G system architecture

The PDU session anchor UPF (i.e. UPF interfacing data network at N6 reference point) acts as the multicast router / switch of corresponding IP PDU session type. The design principle of N6 reference point is to introduce multicast IP datagrams in ingress traffic and thus avoid a need for a dedicated interface to receive multicast traffic in 5G network. The UPF can learn which multicast sources it is serving and what multicast addresses are used, for example via configuration (e.g. from application function). UPF uses a multicast routing protocol to ensure that multicast packets are delivered from neighbouring routers to RAN and the RAN delivers the multicast data to all listening UEs.

The 5G multicast services should be available in dynamic areas where the number of users during popular events (for e.g., in stadiums) can be high and the user distribution within the multicast area will change over the time. In the context of multicast, the RAN is aware of UE's interest to receive data from IP multicast group. Dynamic RAN Multicast Areas (RMA) with synchronization point in NG-RAN can support multitude of deployments from a single cell DU to multiple cells under several DUs, still controlled by a single CU. In Figure 3-2, the Intra-CU cases 1. to 3. are managed with single gNB forming the RMA for IP Multicast transmission, while the case 4. reflects the inter-gNB deployment and for example terrestrial broadcast deployment:

1. Single-Cell PTM, Intra-DU, Intra-gNB

2. Multi-Cell PTM, Intra-DU, Intra-gNB
3. Multi-Cell PTM, Inter-DU, Intra-gNB and
4. Multi-Cell PTM, Inter-gNB

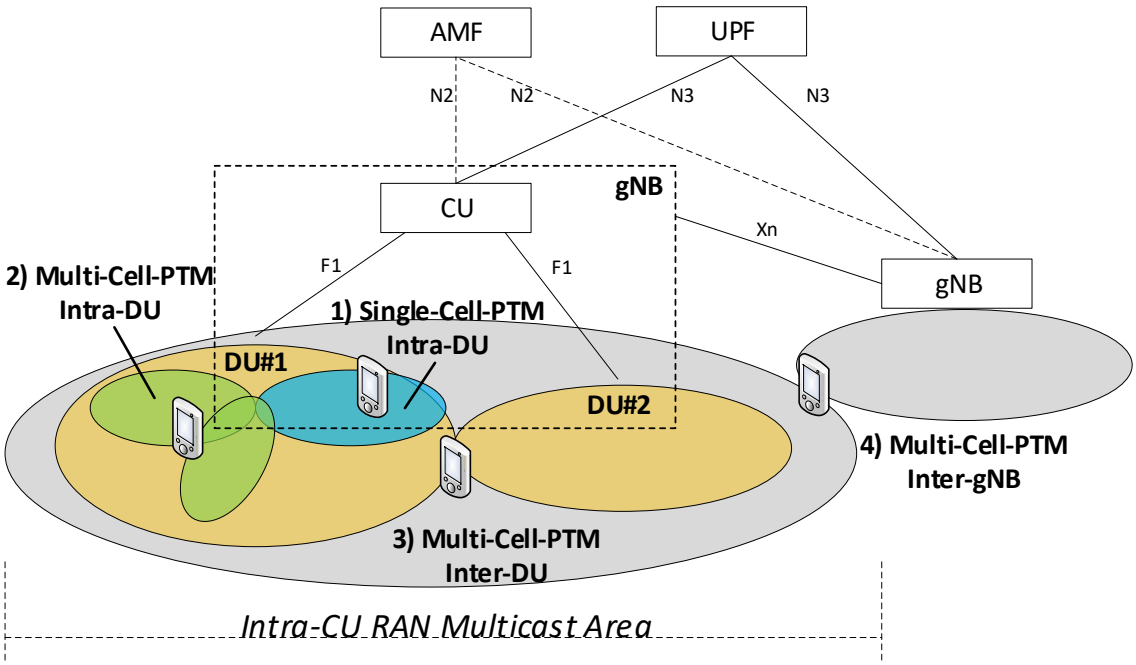


Figure 3-2: Deployments of the RAN Multicast Area, in function of single gNB

3.2 RAN Synchronization for network entities

3.2.1 Overview

The proposed 5G-Xcast RAN architecture does not include a dedicated network entity, which functionality would include the configuration of multi-cell transmission. Instead, the approach uses run-time configuration of the transmission parameters. In the multi-cell transmission, the transmitting gNB-DUs must be synchronized. The gNB-DUs exchange the information about their PHY synchronization/clock and system reference frame number, if this information is not readily available. The PHY synchronization and reference clock information could indicate a synchronization region such as MBSFN synchronization area ID in eMBMS. The gNB-DUs can also determine whether they are synchronized to a common time reference (e.g. GNSS) and provide the PHY synchronization/clock information and system frame number as an offset to the common time reference. The latter approach does not require additional configuration between gNB-DUs.

3.2.2 Synchronized multicast content transmission over multiple cells

Synchronized content delivery over multiple cells by gNB-DUs needs a point for transmission coordination, which is hierarchically above the gNB-DUs in the RAN architecture, e.g. in gNB-CU. The multicast functionality within gNB-CU may be called

gNB-CU-MC. In gNB-CU-MC, the MC stands for Multicast and covers the user plane functions to deliver the multicast data to one or more DUs joining the multicast transmission within the RMA (see Section 5.2, RAN Multicast Area procedures). gNB-CU-MC enforces the synchronized over the air transmission of multicast/broadcast traffic within the gNB-DUs, while fulfilling the QoS targets defined for the traffic flows. One architecture option is to enhance the functions of gNB-CU-CP with multicast operation, to configure gNB-CU-MC and respective gNB-DUs for multicast transmission, so the gNB-CU-MC would cover only the user plane functions. The other option is to extend the gNB-CU-MC with control plane functions so that the gNB-CU-MC covers both the user and control plane functions. Depending on the requirements of deployment, the gNB-CU-MC can be implemented as a stand-alone network entity or integrated into the other functions of gNB-CU, such as gNB-CU-CP or gNB-CU-UP. In this document the gNB-UC-MC is presented as a separate network entity with associated Application Protocol F1-M for clarity of the 5G-Xcast RAN architecture. This approach is presented in Figure 3-3.

Referring to Figure 1-1, the multi-cast enabled gNB uses the SDAP layer where the gNB-CU-MC will receive the multicast / broadcast traffic and identify the type of traffic based on the QoS flow identifier and/or through multicast context associated with the flow. For multicast traffic, the gNB-CU-CP can dynamically configure (add, modify, remove) the gNB-DUs within the multi-cell point-to-multipoint (MC-PTM) list based on the neighbour cell measurement reports from the UEs receiving the traffic. The MC-PTM Setup may contain a request through F1-C to disable link adaptation (LA), report TX parameters for the multicast transport channel (e.g., including the modulation and coding scheme (MCS), TX mode, multiple input multiple output (MIMO) configuration, downlink coordination information – which could also include physical resource block (PRB): fixed or list of starting PRBs for PDUs with sequence numbers.

For broadcast traffic, the SDAP can map the flows into broadcasted data radio bearer or specific multicast radio bearer with support for signalling information for broadcast. Flows are broadcasted using the broadcast channels by the appropriate multi-cell setup of gNB- determined based on the RAN Multicast Area where the data needs to be multicasted or broadcasted.

The gNB-CU-CP configures gNB-CU-MC to trigger F1-C or F1-M interface setup for gNB-DUs which are in the MC-PTM transmission list via the E1-C interface. Here F1-M represents a logical interface between gNB-CU entity and gNB-DU entity, where the interface can be a unicast or a multicast tunnel of multicast radio bearer, e.g. GTP-U over IP multicast as used in M1 interface of eMBMS.

For content synchronization the F1-U interface could be modified to include a new RAN-SYNC protocol enhancement.

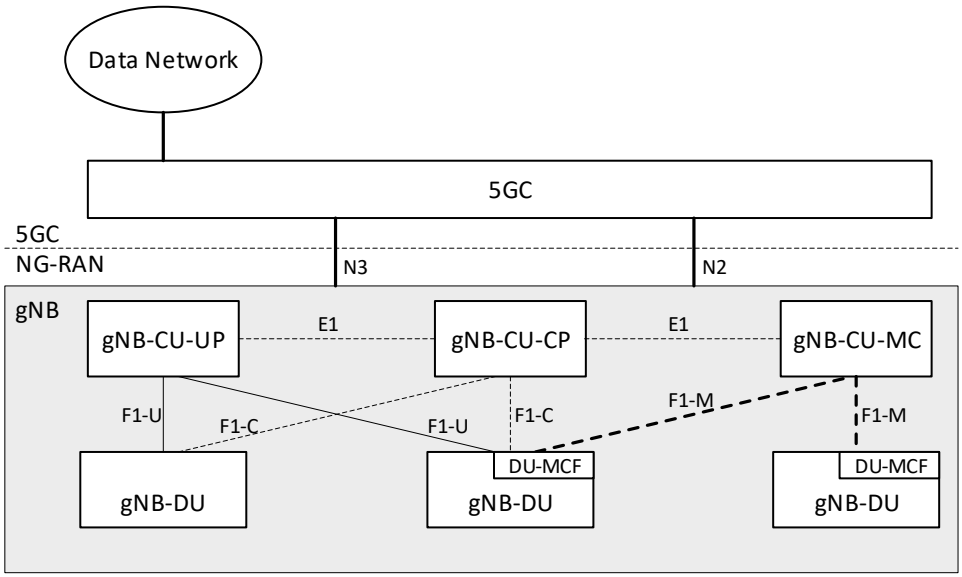


Figure 3-3: NG-RAN architecture with RAN-based synchronization of multicast / broadcast traffic

3.2.3 Synchronized multicast content transmission over Single Frequency Networks

Single Frequency Networks (SFN) is a technique of CoMP (Coordinated MultiPoint) which improves coverage at cell edge by exploiting constructive interference. SFN networks are widely used in state-of-the-art Terrestrial Broadcast deployments. In order to deploy an SFN, two main requirements must be met: Common use of radio parameters between all transmitters in the SFN service area and common scheduling of data at PRB level. Currently in Rel-15, the functionality to perform common scheduling and MCS between different cells is not standardized. Fortunately, other radio parameters such as cyclic prefix, scrambling and DMRS reference signal coding can be configured in Rel-15. More details about the 5G-Xcast air interface can be found in D3.2 [3].

For the same multicast data traffic delivery, the concept of RAN Multicast Area (RMA) is defined and presented in Chapter 5.2. Two metrics were chosen inside 5G-Xcast to define the appropriate RMA deployment: the number of users and their geographical location. Generally, for ensuring synchronous transmissions, the same MCS and the same assignment to PRBs must be followed by every transmitter forming the RMA. Depending on which cells are included in the RMA, different synchronization scenarios will occur, and the SFN requirements must be supported in the appropriate interfaces. This in mind, three different SFN scenarios are identified to deploy SFN within RMA. To perform an analysis of the synchronization options, we assumed that if the cells or RRH served by a gNB-DU are geographically close (less than 1.5 Km) then there is no noticeable network delay.

Also, if the network delays experienced by the transmitters involved in the PTM transmissions are different (> 5 μs according to 3GPP TS 36.133), SFN operation is not possible without additional help of a protocol that compensates these delays. In the case of MBMS, this protocol is called SYNC [8] and it is applied by the BM-SC in the Core Network, encapsulating the data with Time-to-Air headers.

In 5G-Xcast, to ensure the synchronization requirements, a revision of the SYNC protocol is proposed, named RAN-SYNC. In eMBMS, a mobile operator had to properly configure SYNC parameters in both the BM-SC and each of the transmitters involved in the MBSFN transmission. The lack of channel feedback of propagation conditions and QoS from the eNB towards the core network hindered the troubleshooting of the service delivery problem and adaptation of the MBSFN transmission parameters to better meet the target QoS. To solve this, RAN-SYNC protocol is applied from the anchor gNB, reaching all the cells forming the SFN. Note that, the option to use eMBMS SYNC is still available since the XUF [1] implements this functionality.

RAN-SYNC protocol is divided into Control Plane and User Plane, connected via E1 interface. The Control Plane functionality is implemented inside the gNB-CU-CP of the anchor gNB and takes care of the initial provisioning of the SYNC parameters when establishing the broadcast flow or in the multicast PDU context modification procedure. By introducing this, the SFN configuration is easier and faster compared to eMBMS. Also, the RAN-SYNC tracks in real-time the transmission delays between the group of transmitters in order to generate a proper Time-to-Air timestamps. For example, the Time-to-Air generated by the gNB-CU-CP in a given SFN can be the largest propagation delay plus a small buffer margin. On the other hand, the User Plane encapsulates the multicast or broadcast data with SYNC headers. Existing headers in eMBMS SYNC [8] are reused. This functionality resides in the gNB-CU-MC. An example of RAN-SYNC can be seen in Figure 3-4.

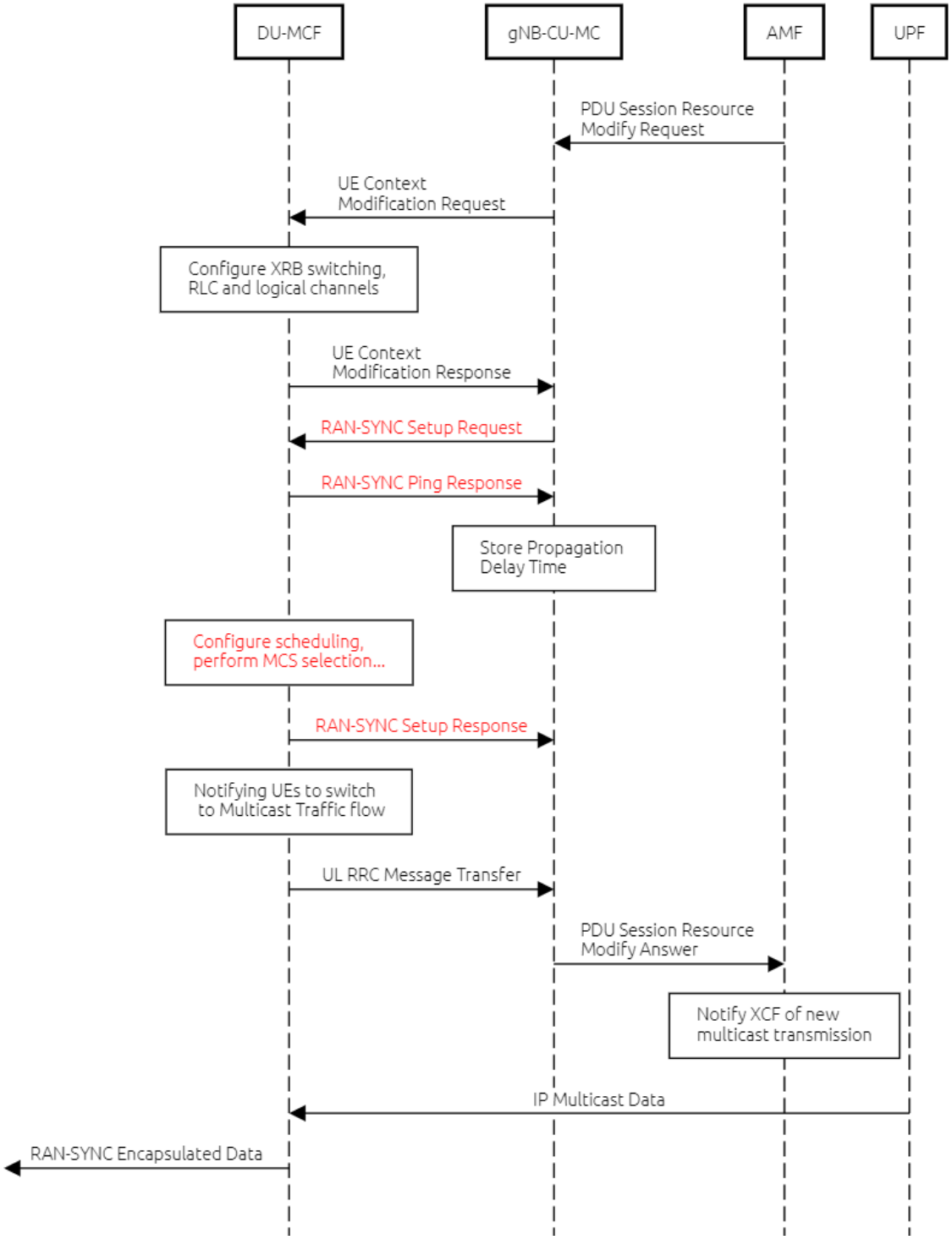


Figure 3-4: RAN-SYNC procedure applied in Transparent Multicast mode in a Cloud-RAN deployment.

Depending on the cells forming the SFN, several synchronization scenarios are possible. In the first synchronization scenario (Figure 3-5) occurs when every cell involved in the multicast transmission is within the same gNB-DU. In this scenario, the DU-MCF (Distributed Unit-Multicast Function) module enforces the same scheduling at MAC level. The second SFN synchronization scenario in Figure 3-5 considers cells belonging to two or several gNB-DUs governed by the same gNB-CU to form the RMA.

SFN requirements must be fulfilled through F1 interface, including resource scheduling and the setup of radio parameters (Numerology and MCS) for every transmitter. For both scenarios, the decision to apply RAN-SYNC will depend on the propagation delay measured by the gNB-CU-MC.

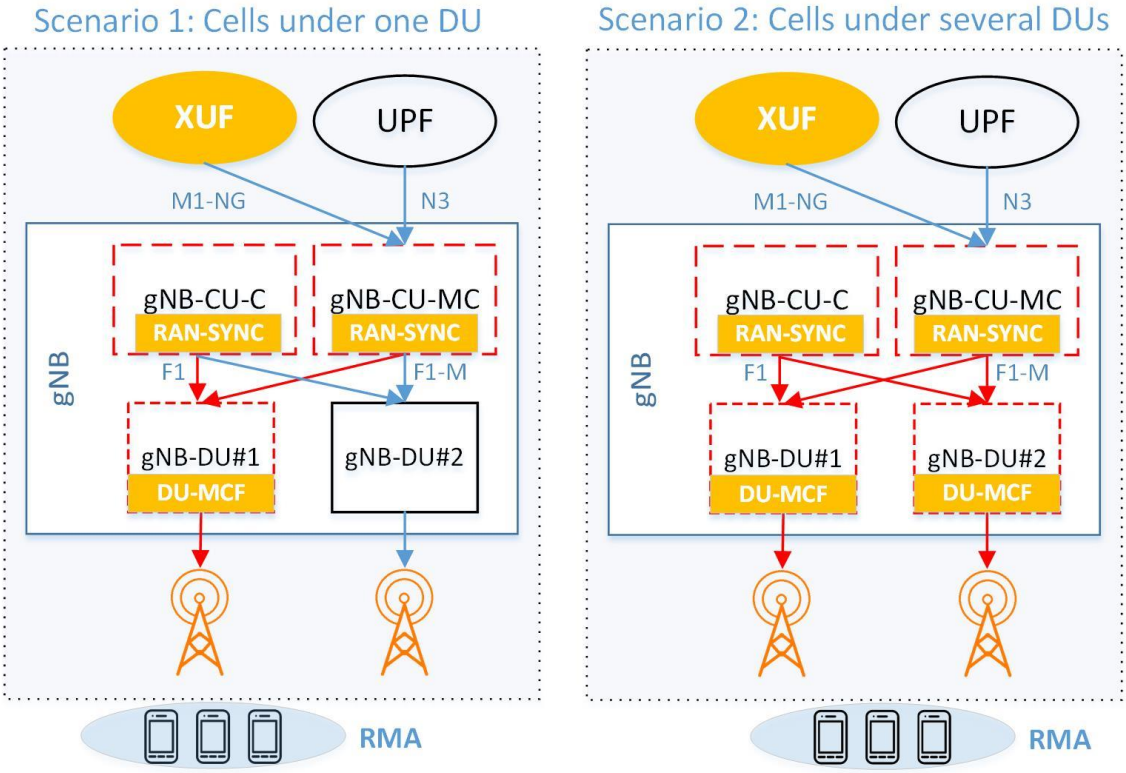


Figure 3-5: Synchronization scenarios for one or multiple gNB-DUs

Lastly, for the third SFN synchronization scenario in Figure 3-6, the relevant RMA is formed by cells belonging to two or more gNB-CUs or gNBs. The synchronization needs to be performed through an enhanced Xn interface, where RAN-SYNC carries synchronization parameters, and additional signalling conveys the MAC scheduling and radio parameters. Since several gNBs could be involved in the SFN, additional requirements arise. The anchor gNB-CU decides the radio parameters and the scheduling, and anchor propagates that decision to the secondary gNB-CUs to minimize latency, while the secondary gNB-CU have their RAN-SYNC Control module deactivated for this SFN. Even with this coordination method, the SFN signalling must go through Xn and F1 interface for the secondary gNBs. The jitter introduced by Xn can hinder the SFN operation, so it is required that the anchor gNB-CU-MC has propagation delay information of the other gNB cells. Table 1 summarizes all the SFN scenarios.

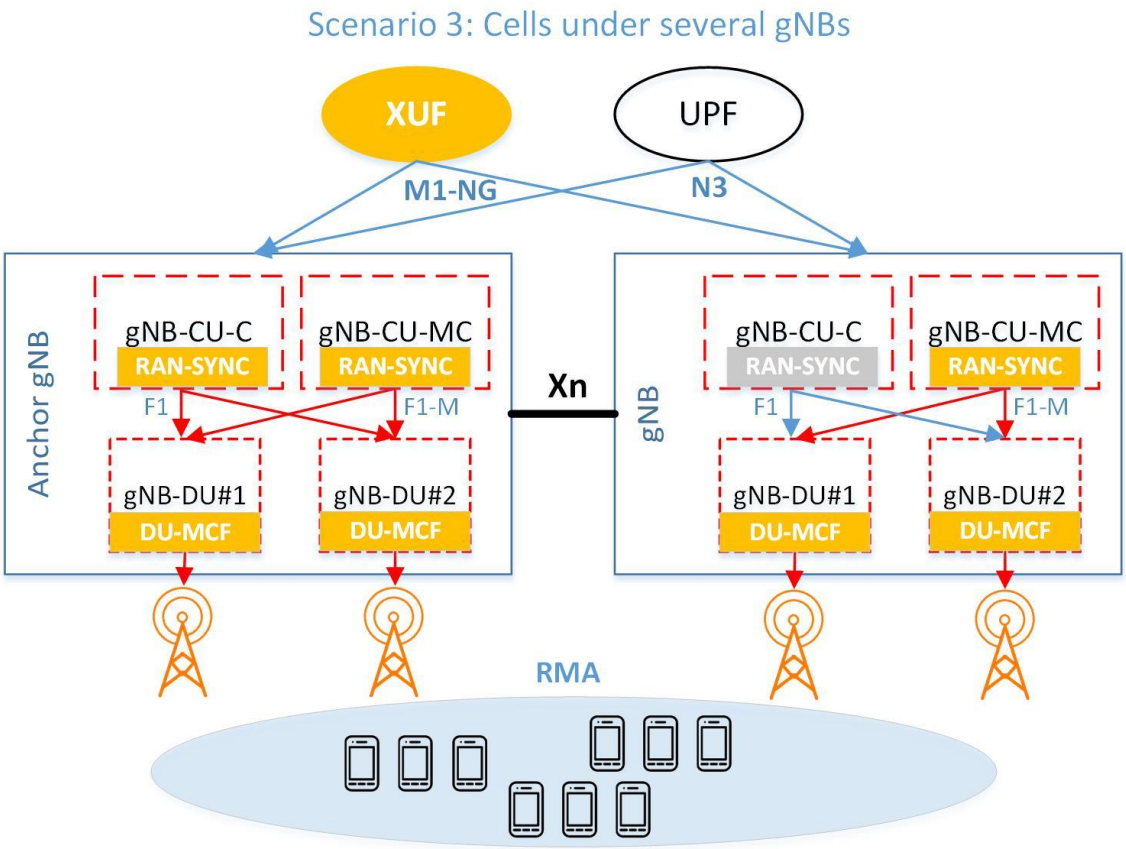


Figure 3-6: Synchronization scenario for multiple gNB-CUs

Scenario	Interfaces involved	Notes
Intra-DU	F1, F1-M	gNB-CU-CP or gNB-CU-MC signals the scheduling, numerology and radio parameters to DU-MCF via F1-M interface.
Inter-DU	F1, F1-M	gNB-CU-CP or gNB-CU-MC signals the scheduling, numerology and radio parameters to DU-MCFs via F1-M interface.
Inter-CU	F1, F1-M, Xn	Anchor gNB-CU-MC signals the scheduling, numerology and radio parameters via Xn interface to other gNB and F1-M for its cells.

Table 1: Comparison of synchronization scenarios

3.2.4 Dimensioning limits of the content synchronization area using single gNB

Content synchronized MBSFN transmission in LTE requires the MCE, as well as the SYNC time synchronization protocol running between eNB and BM-SC. The

Synchronization protocol is an additional layer above the transport protocol to synchronize the delivery of eMBMS content from multiple eNBs by reordering and detecting lost SYNC packets. The MBSFN area as defined in [8], is an area in which cells transmit the same multimedia content with same radio configuration, at the same time, controlled by a single MCE.

In 5G-Xcast RAN architecture, the proposed content synchronization solution uses RAN-based time synchronization hosted in gNB-CU-MC, i.e. synchronization within one gNB considering the higher layer split (Chapter 6.1, 5G-Xcast Deployment Scenarios) and the CP and UP split of gNB. In this solution, the proposed 5G-Xcast RAN architecture would need to support a deployment scenario with synchronous transmission by cells belonging to one gNB in a large SFN area. When the content synchronization is performed only in RAN under one gNB, the limiting factor should not be the number of cells which a single gNB can support, as the following figures show.

According to 3GPP TS 38.473 [35], the dimensioning limits with one gNB deployment are:

1. gNB-DU ID is an integer with range from 0 to 2^{36}-1, where the gNB-DU ID uniquely identifies the gNB-DU at least within a gNB-CU. The gNB-DU ID is independently configured from cell identifiers, i.e. there is no connection between gNB-DU ID and cell identifiers.
2. Maximum number of cells that can be served by a gNB-DU is 512.
3. NR CGI = PLMN + 36-bit string
4. gNB-CU UE F1AP ID and gNB-DU UE F1AP ID are integers in range from 0 to 2^{32} -1, where the gNB-CU UE F1AP ID and gNB-DU UE F1AP ID uniquely identify the UE association over the F1 interface within the gNB-CU and gNB-DU respectively.

Therefore, it can be concluded that the addressing ranges of relevant RAN node identifiers can support practical nation-wide deployment with a single gNB. It is not in the scope of this deliverable if the approach of gNB-CU-MC in one gNB would be a preferred implementation option nation-wide deployment.

3.2.5 Summary

Considering the section 3.2.2 and Cloud-RAN and centralized deployment of gNB-CU-CP in section 6.1.3, Cloud-RAN based deployment, gNB-DU can be connected to any suitable gNB-CU-CP. When deploying SFN all gNB-DUs controlling cells in the given geographical area could be connected to one gNB-CU-CP. One valid deployment option is the scenario where one gNB-CU-MC and one or more gNB-CU-UPs are providing content synchronization and multicast traffic delivery with the described RAN-based synchronization solution controlled by gNB-CU-CP.

Thus, a large SFN network, such as nation-wide area, can be supported by RAN-based synchronization solution where

- Scenario 1: synchronous transmission by cells under the control of multiple gNB-CU-CPs with shared gNB-CU-MC
- Scenario 2: synchronous transmission by cells under the control of one gNB-CU-CP with shared or dedicated gNB-CU-MC.

3.3 Summary

In this section, the 5G-Xcast RAN Logical Architecture is presented. This architecture supports both multicast and broadcast transmissions, and it is based on concepts of

RMA, RAN based synchronization and the gNB-CU-MC network entity to deliver the multicast and broadcast traffic to users. RMA is a coordinated group of cells forming an area where the PTP, PTM, MC-PTM or SFN transmission modes can be selected. The RMA is managed by the gNB-CU-CP at control plane and gNB-CU-MC for the user plane. In order to account to the transmission delays between different cells, a revised version of compensation protocol SYNC, named RAN-SYNC, is introduced. This protocol enforces the synchronization parameters between all the relevant cells involved in the RMA, facilitating and improving the deployment operation, compared to eMBMS SYNC. Finally, the address space of network identifier entities and related interfaces for the signalling is shown to enable one gNB managing a large area, e.g. nation-wide, SFN network.

4 5G-Xcast RAN Interfaces

4.1 Introduction

ITU-T Recommendation I.112 [5] defines an interface as “the common boundary between two associated systems”, and 3GPP follows the definition from 3GPP TR 21.905 [6]. A network interface covers all protocol layers of significance for network elements at both sides of the interface. E.g. if the network elements are L2 entities then the interface should be clearly specified at L2. In most cases, the interface specification goes all the way down to L1, to be represented as a full protocol stack to enable the interconnection and even plug and play. RAN interfaces are categorised into external and internal interfaces. The external interfaces are those between the 5G RAN (named NG-RAN in 3GPP) and the 5GC, and those between the 5G RAN and the UE. The internal interfaces are those between 5G RAN nodes. 3GPP has been continuously working on the definition and standardisation of those interfaces in the 5G system. The principles on the 5G-Xcast RAN interface aspects are:

- Endeavour to reuse and to enhance current NG-RAN interfaces to support broadcast and multicast to keep the system interfaces as simple and as few as possible.
- Define new interfaces to support broadcast and multicast if it is necessary.

The network interfaces should allow easy interconnection of products from different vendors, and the possibility of forward compatibility for future evolution.

4.2 Interfaces between 5G-Xcast RAN and 5GC

3GPP has defined the interface between the 5G RAN and 5GC as NG, and further specified into NG-C and NG-U for CP and UP separately [11][18][19][20][21][22]. NG-C maps to the reference point N2 and NG-U to the reference point N3 [7]. Specifically, as shown in Figure 4-1, N2 marks the interface between a gNB and the AMF, and N3 marks the interface between a gNB and the UPF. In 5G-Xcast, in order to support the system architecture alternative 2 (see Annex A.1), where XUF is directly connected to the RAN, a new UP interface M1-NG is introduced, marking the interface between the broadcast and multicast supporting gNB and the XUF. M1-NG is optional and is needed only for system architecture alternative 2.

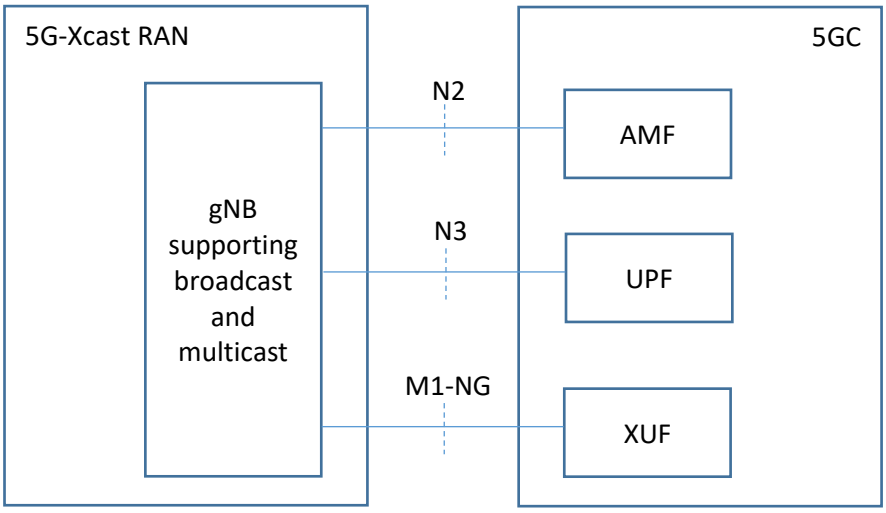


Figure 4-1: 5G-Xcast RAN and 5GC interfaces in 5G-Xcast reference architecture

4.2.1 Interface N2

The N2 protocol stack is shown as in Figure 4-2 [18]. The Transport Network Layer (TNL) is built on IP [41] transport. SCTP [42] is used for the transport of the application layer signalling NGAP (NG Application Protocol). In 4G LTE, to support broadcast and multicast, a CP interface M3 [9] is defined between the E-UTRAN and EPC, specifically between the MCE and the MME [8]. In 5G-Xcast RAN, similar functionality to eMBMS MCE is handled by gNB-CU-CP (section 3.2.2) and integrated with gNB, resembling the eMBMS distributed deployment architecture in LTE. The other option is to extend the gNB-CU-MC with control plane functions so that the gNB-CU-MC covers both the user and control plane functions. Broadcast and multicast are supported by the enhanced AMF in the 5G-Xcast core network [1]. N2 is enhanced to fully support the functions of the M3 interface in LTE, which means NGAP also accordingly is enhanced with the M3AP functionalities.

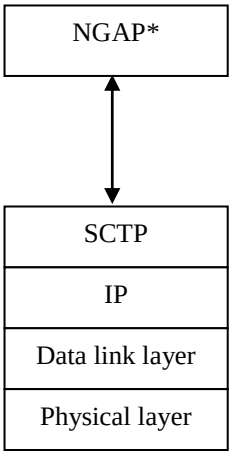


Figure 4-2: N2 protocol stack between gNB and AMF [18]

NGAP*: enhanced to support broadcast and multicast

The 5G-Xcast N2 functions include all the control functions defined for NG interface in [18], and additionally, to support broadcast and multicast, include the followings:

- Enhanced PDU Session Management Function to handle broadcast and multicast sessions;
- Enhanced N2 Configuration Function to configure gNB-CU-MC and related multicast coordination and configuration functions integrated with the gNB.

Signalling procedures for N2 are extended to support broadcast and multicast, including:

- Broadcast and multicast session management procedures;
- Broadcast and multicast UE context management procedures;
- Broadcast and multicast UE mobility management procedures;
- AMF/XCF management procedures.

4.2.2 Interface N3

The N3 protocol stack is shown as in Figure 4-3, as specified in [18]. The TNL is built on IP transport and IP Multicast is used for point-to-multipoint delivery of user packets. GTP-U [39] upon UDP [40] provides non-guaranteed delivery of UP PDUs between the

gNB and the UPF. N3 fully supports the functions of the M1 interface in LTE, in the cases of 5G-Xcast architecture Alternatives 1 and 3, where broadcast and multicast UP data will be carried over N3 between gNB and UPF. On top of TNL, unicast, multicast and broadcast UP PDU Sessions are multiplexed at RNL. SYNC protocol [8] interface is supported at the PDU Session layer if Alternative 1 or 3 is the deployed architecture.

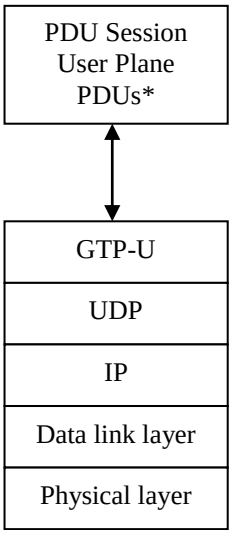


Figure 4-3: N3 protocol stack between gNB and UPF [18]

PDU Sessions*: including broadcast and multicast sessions over SYNC

The following functions are supported over N3, as specified in [8]:

- Transfer of user data alongside synchronisation information;
- Transfer of synchronisation information without user data.
- Enhanced N3 user plane function to manage received multicast traffic and deliver the traffic to DUs configured to join the multicast transmission.

4.2.3 Interface M1-NG

The M1-NG interface is introduced particularly for the 5G-Xcast architecture Alternative 2, where the XUF is connected directly to the gNB. M1-NG is defined same as M1 in LTE, and also as N3, using the same TNL structure as shown in Figure 4-4. The identical TNL allows the maximum compatibility in implementation, allowing efficient network deployment and component reuse. The SYNC interface is also provided between the gNB and the XUF, on top of the TNL.

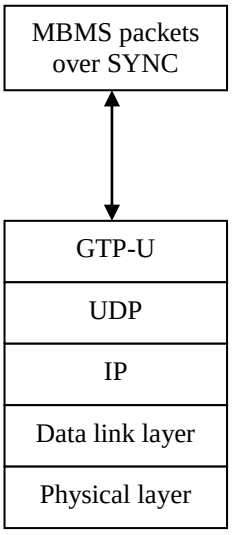


Figure 4-4: M1-NG protocol stack between gNB and XUF

4.3 5G-Xcast RAN Internal Interfaces

The 5G RAN is providing CU-DU split as specified in 3GPP (NG-RAN [17]). The gNB functions are split into CU and DU, where DU entails protocol functions of PHY, MAC and RLC, and CU entails all other higher layer functions from PDCP and above. F1 interface [32][33][34][35][36] is defined between CU and DU. In a typical deployment, CUs are put into a pool and form the cloud. CUs are interconnected through Xn interface [23][24][25][26][27]. In 5G UP and CP are clearly separated, and consequently F1 is separated into F1-C on CP and F1-U on UP, Xn into Xn-C on CP and Xn-U on UP. A gNB-CU is further separated logically into gNB-CU-CP on CP, gNB-CU-UP on UP, and gNB-CU-MC connected with the interface E1 [28][29][30][31], where gNB-CU-MC is introduced to support broadcast and multicast, and a new interface F1-M is introduced to connect it with the gNB-DU. In order to support wireless relay [37][38], the NG-RAN Un interfaces on both CP and UP are introduced as Un-C and Un-U, to connect the IAB nodes. The interfaces within the reference architecture are shown as in Figure 4-5.

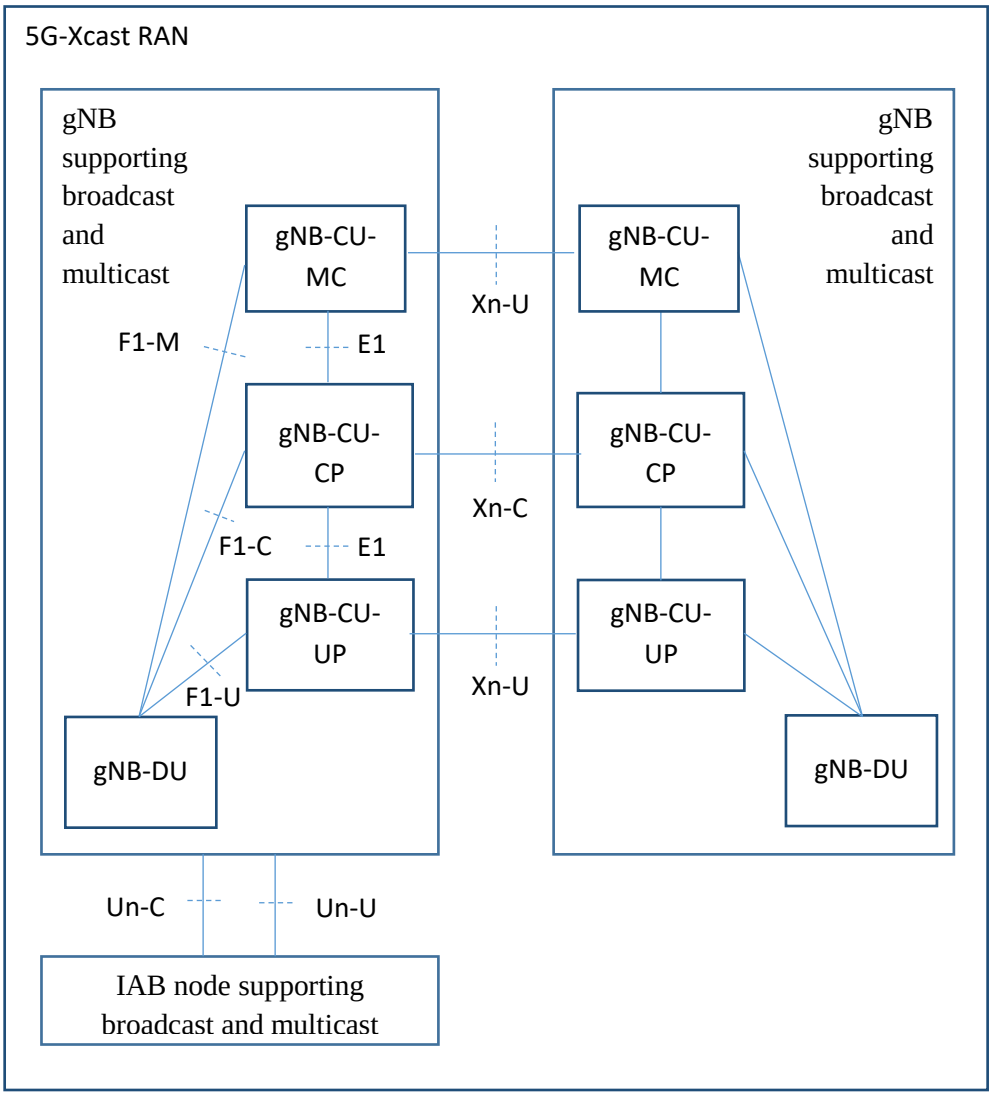


Figure 4-5: Internal interfaces in 5G-Xcast RAN reference architecture

In this section we exhaustively examine each one of the internal interfaces, illustrating their protocol structure, outlining the extended functions and procedures in order to support broadcast and multicast.

4.3.1 Interface F1-C

Figure 4-6 shows the protocol structure for F1 on Control Plane (F1-C) as specified in [32] . The TNL is based on IP transport. SCTP is used to transport the application layer signalling protocol F1AP (F1 Application Protocol) on RNL.

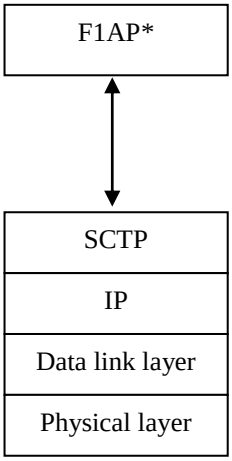


Figure 4-6: F1-C protocol stack between gNB-CU and gNB-DU [32]

F1AP*: enhanced to support broadcast and multicast

In 5G-Xcast the F1-C functions are extended with followings to support broadcast and multicast:

- F1 interface management function, including F1-M management;
- Broadcast and multicast related System Information management function;
- F1 broadcast and multicast UE context management function;
- Broadcast and multicast related RRC message transfer function.
- Enhanced N2 Configuration Function to manage MC integrated with the gNB.

Signalling procedures for F1-C are extended with followings to support broadcast and multicast, including:

- F1 management procedures, including F1-C, F1-U, and F1-M;
- Broadcast and multicast UE context management procedures;
- Broadcast and multicast related RRC message transfer procedures;
- Broadcast and multicast related System Information procedures.

4.3.2 Interface F1-U

Figure 4-7 shows the protocol structure for F1 on User Plane (F1-U) as specified in [32]. The TNL is based on IP transport. GTP-U on top of UDP is used to transport the mixed UP PDUs of unicast, multicast and broadcast.

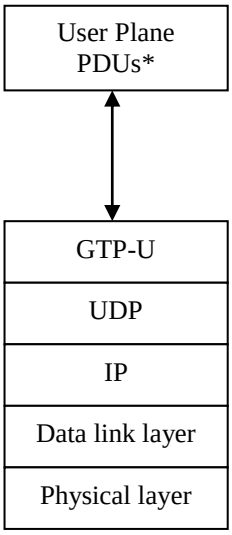


Figure 4-7: F1-U protocol stack between gNB-CU and gNB-DU [32]

User Plane PDUs*: including broadcast and multicast PDUs

In 5G-Xcast the interface supports IP Multicast at TNL to transfer broadcast and multicast RLC frames at RNL. Extended from [32], F1-U functions include:

- Transfer of user data, including unicast, multicast and broadcast data;
- Broadcast and multicast flow control function.

4.3.3 Interface F1-M

Figure 4-8 shows the protocol structure for F1 on Multicast (F1-M) as proposed in section 3.2.2 Synchronized multicast content transmission. The F1-M tunnel can be unicast or multicast tunnel, e.g. GTP-U over IP multicast. In the case that the tunnel is a unicast tunnel, the tunnel may be modified to a multicast tunnel. The target gNB-DU sets up the F1-M tunnel. In the case when F1-M tunnel uses unicast tunnelling, the target gNB-DU needs to allocate DL tunnel end-point and provide the tunnel information to the gNB-CU-CP.

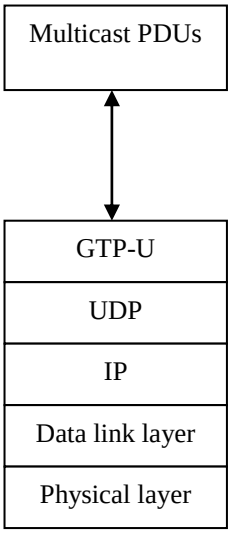


Figure 4-8: F1-M protocol stack between gNB-CU and gNB-DU

4.3.4 Interface E1

Figure 4-9 shows the protocol structure for E1 as specified in [28][29][30][31]. The TNL is based on IP transport. SCTP is used to transport the application layer signalling protocol E1AP (E1 Application Protocol).

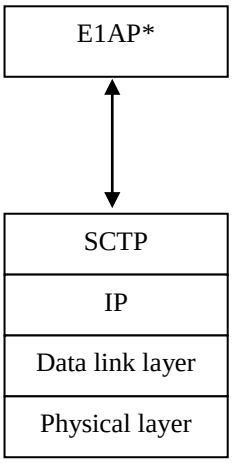


Figure 4-9: E1 protocol stack between gNB-CU-CP and gNB-CU-UP [28]

E1AP*: enhanced to support broadcast and multicast

Apart from the E1 functions specified in [32], the enhancement includes similar functions as provided by M2 interface in E-UTRA [9]:

- Broadcast and multicast session handling function: broadcast and multicast session start, stop and update;
- Broadcast and multicast scheduling information provision function;
- Broadcast and multicast service counting function;
- Broadcast and multicast suspension and resumption function.

E1 procedures are extended with followings to support broadcast and multicast:

- E1 management procedures to support connection to gNB-CU-MC;
- Broadcast and multicast bearer context management procedures.

4.3.5 Interface Xn-C

Figure 4-10 shows the protocol structure for Xn on Control Plane (Xn-C) as specified in [23]. The TNL is based on IP transport. SCTP is used to transport the application layer signalling protocol XnAP (Xn Application Protocol).

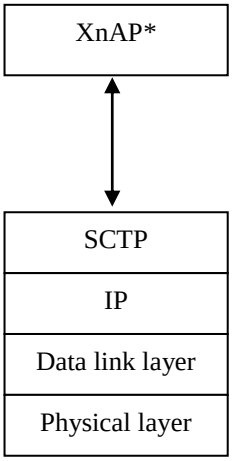


Figure 4-10: Xn-C protocol stack between gNB-CU-CPs [23]

XnAP*: enhanced to support broadcast and multicast

In 5G-Xcast the Xn-C functions are extended with followings to support broadcast and multicast:

- Xn interface management to support interconnection of gNB-CU-MCs;
- Broadcast and multicast UE mobility management functions;
- Broadcast and multicast related Dual Connectivity;
- Broadcast and multicast related resource coordination function.

Signalling procedures for Xn-C are extended with followings to support broadcast and multicast, including:

- Interface management procedures to support interconnection of gNB-CU-MC;
- Broadcast and multicast UE mobility management procedures;
- Broadcast and multicast related Dual Connectivity procedures;
- Broadcast and multicast related resource coordination procedures.

XnAP is enhanced to coordinate the gNB-CU-MC functionality in the centralised gNB-CU pool such that the broadcast and multicast transmission can be coordinated among the cells of different gNB in a same broadcast and multicast service area, especially for adjacent gNBs for service continuity in the roaming scenario.

4.3.6 Interface Xn-U

The Xn on User Plane interface (Xn-U) is defined between two NR RAN nodes as specified in [23]. Figure 4-11 shows the protocol structure for Xn-U. The TNL is based on IP transport. GTP-U on top of UDP is used to transport the mixed UP PDUs of unicast, multicast and broadcast between two NR RAN nodes.

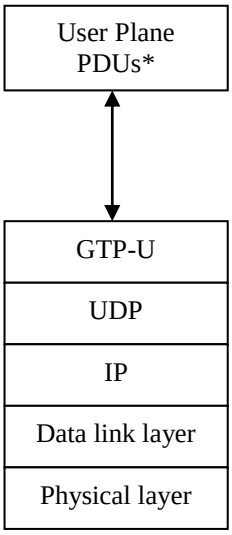


Figure 4-11: Xn-U protocol stack between gNB-CU-UPs [23]

User Plane PDUs*: including broadcast and multicast PDUs

In 5G-Xcast the Xn-U functions are extended with followings to support broadcast and multicast:

- Data transfer function, including unicast, multicast and broadcast data;
- Broadcast and multicast flow control function;
- Broadcast and multicast related assistance information function;
- Broadcast and multicast related fast retransmission function.

Extended from [23], the Xn-U procedures in 5G-Xcast include:

- Transfer of broadcast and multicast service data procedure, including both downlink PDCP data procedure and PDU Session data procedure;
- Broadcast and multicast data delivery status procedure;
- Broadcast and multicast related assistance information procedure.

4.3.7 Interface Un-C

In a relay scenario where the relay node (IAB node) functions as a gNB, control signalling messages are transported over SCTP with enhanced XnAP and NGAP. The transmission utilises NR air interface between the anchor (donor) gNB and the IAB node, and between the IAB nodes. IP packets transporting the signalling messages are mapped to the SRB directly of the NR protocol stack. The TNL structure is shown in Figure 4-12, on the NR Un on Control Plane (Un-C). As can be seen the interface is compatible with Xn-C.

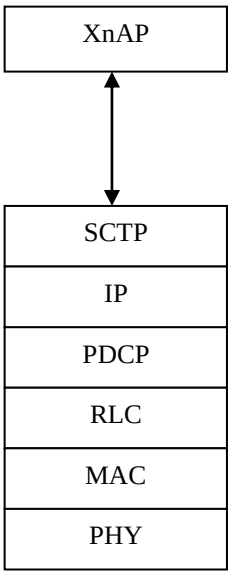


Figure 4-12: Un-C protocol stack between gNB and IAB nodes and that between IAB nodes, where IAB nodes function as gNB

4.3.8 Interface Un-U

UP PDUs between the anchor gNB and IAB nodes, and that between the IAB nodes, are transported over GTP-U and the full NR stack. IP packets transporting the user data are mapped to the SDAP QoS flows and further to DRB of the NR protocol stack. The TNL structure is shown in Figure 4-13, on the NR Un on User Plane (Un-U). As can be seen the interface is compatible with Xn-U.

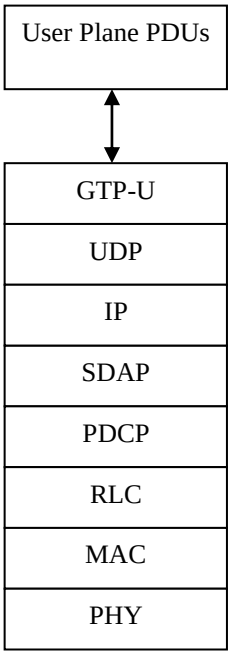


Figure 4-13: Un-U protocol stack between gNB and IAB nodes and between IAB nodes, where IAB nodes function as gNB

4.4 Interface between 5G-Xcast RAN and UE

Uu is the air interface between the 5G-Xcast RAN and the UE, as shown in Figure 4-14.

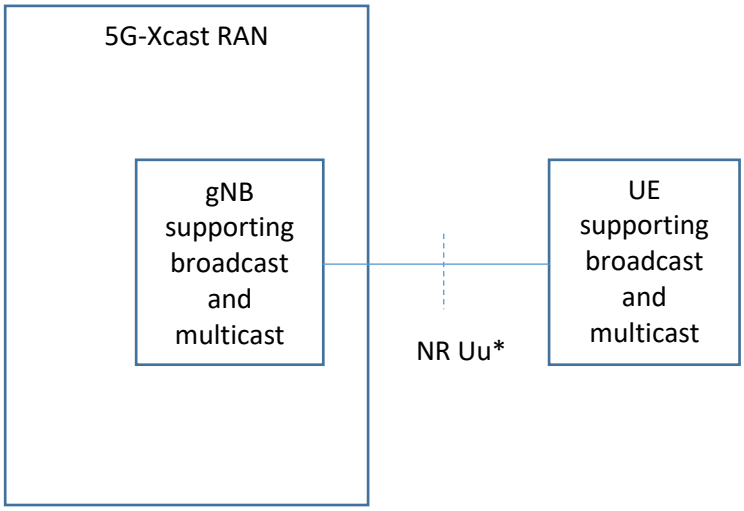


Figure 4-14: Uu interface in the 5G-Xcast reference architecture

NR Uu*: enhanced to support broadcast and multicast

4.4.1 Interface Uu

RRM and RRC between the gNB and the UE are performed via the CP protocol stack as in Figure 4-15, which enhanced the stack in [11] and specified in appendix A.2.

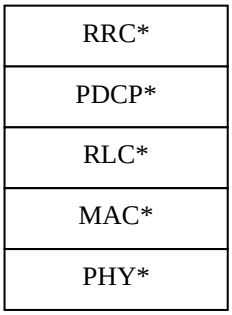


Figure 4-15: Uu protocol stack on CP

RRC*/PDCP*/RLC*/MAC*/PHY*: enhanced to support broadcast and multicast

In order to support broadcast and multicast in the 5G network, a new logical channel of MTCH is introduced. The protocol stack is compatible with [11] and specified in appendix A.2. The UP protocol stack is shown in Figure 4-16.

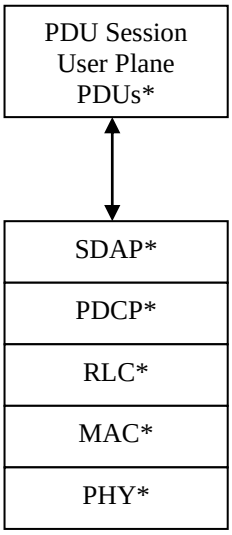


Figure 4-16: Uu protocol stack on UP

User Plane PDUs*: including broadcast and multicast PDUs

SDAP*/PDCP*/RLC*/MAC*/PHY*: enhanced to support broadcast and multicast

4.5 Summary

This section described the interfaces needed to deploy 5G-Xcast RAN. The described network interfaces also cover the protocol layers needed to introduce the connectivity between network entities, and between 5G-Xcast RAN and UE. The new interfaces are needed to implement the broadcast and multicast functions. The M1-NG interface between 5G-Xcast RAN and 5GC was introduced for the 5G-Xcast architecture Alternative 2 to connect XUF directly to the gNB. The new 5G-Xcast RAN internal interface F1-M interface between gNB-CU-MC and DU enables the gNB-CU-MC entity to deliver the broadcast and multicast traffic in RAN to transmitting DUs.

The design was following the principle where the reuse of Rel-15 5GC and NG-RAN interfaces was maximized and new interfaces to support broadcast and multicast if it is necessary.

5 5G-Xcast Protocol Architecture and Procedures

5.1 Introduction

The 5G-Xcast services and vertical segments set a variety of very diverse requirements. The design of 5G-Xcast RAN protocol architecture and procedures should consider the design principles where the multi-service RAN architecture needs to be flexible and support the coexistence of PTP, SC-PTM, MC-PTM and broadcast transmissions. Baseline for the RAN logical architecture design is NG-RAN Release 15 architecture.

To allow deployment of existing Multicast/Broadcast services and new services, the overall 5G-Xcast RAN architecture needs to support both (i) dynamic adjustment of the Multicast/Broadcast area based on the user distribution or service requirements and (ii) allow static and dynamic resource allocation between unicast and Multicast/Broadcast. Further, the RAN architecture should support full allocation of downlink carrier resources for Multicast/Broadcast in large geographical areas up to the size of an entire country in SFN mode.

The requirement to support both dynamic local multicast operation and the nation-wide SFN modes put a special challenge to RAN architecture design. This section concentrates on the architecture aspects and the higher layer protocol design. The more detailed protocol aspects are addressed in the 5G-Xcast deliverable D3.4.

5.2 RAN Multicast Area procedures

5.2.1 Introduction

One key challenge of IP multicast content delivery in 5G networks is to manage the multicast area, e.g. RAN based Multicast Area (RMA), where the procedures are used to maintain the multicast area along with UE mobility and connectivity. SC-PTM transmission method was developed to support content delivery to a smaller area. The design target of RMA is to enable dynamic areas based on user geographical distribution, reusing the flexibility of the unicast architecture and basic principles of SC-PTM extended over to Multi-Cell PTM (MC-PTM).

Users having active unicast traffic will be in RRC_CONNECTED state [11] and since the UE location is known by a single cell, it is proposed that the RAN (e.g. anchor gNB) will decide the multicast bearer configuration or deliver the multicast traffic over unicast data radio bearers. If the number of active users is low, the multicast traffic is delivered to UEs using unicast.

When the unicast traffic of a UE is detected to have low activity, the UE is moved to RRC_INACTIVE and the UE continues to receive multicast traffic within the configured RMA. The RMA, where the UE can receive multicast traffic, is defined and controlled by RAN and can be part of the RRC configuration, or part of the broadcasted multicast configuration (e.g. System Information). The anchor gNB (usually the last serving gNB) defines the RMA configuration, and in case of multiple gNBs, distributes it over Xn interface to the gNBs which belong to the RMA.

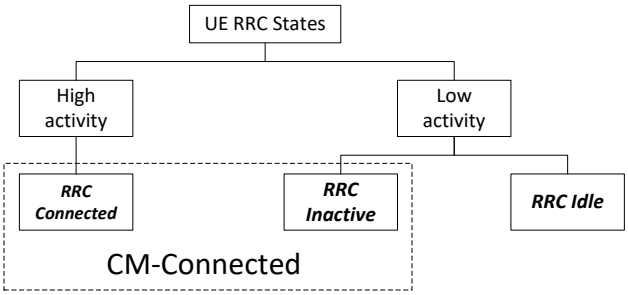


Figure 5-1: Unicast activity and RRC States

Depending on the number of low activity UEs receiving multicast in a cell, the gNB can decide to keep one (or more) UEs in RRC_CONNECTED state, assuming that multicast bearer mapped to unicast bearer or direct usage of unicast bearer will be more spectral efficient than multicast bearer in RRC_INACTIVE state with limited feedback. The benefit of RRC_INACTIVE over RRC_IDLE is the maintained connection to AMF/UPF where the connection management state remains in CM-Connected and the UE Context is stored in both UE and RAN. This will allow low latency state transition between RRC_CONNECTED and RRC_INACTIVE.

The 5G RAN logical architecture with two RMA Ids is presented in Figure 5-2, including three UEs receiving IP multicast traffic. This can be further described with three examples.

1. Based on Figure 5-1, as an example, UE1 may be in RRC_CONNECTED state receiving both unicast and PTM multicast traffic from the same DU. Location of UE1 is known by a single cell in RAN, thus enabling the transmission of unicast and multicast traffic using only unicast bearers.
2. The UE2 has completed its unicast traffic, and due to low activity, the RAN (e.g. anchor gNB) decides to suspend the RRC configuration and configures the UE2 into RRC_INACTIVE state. The multicast traffic will move from unicast Radio Bearer (RB) to multicast RB thus allowing the UE2 to continue the reception of PTM multicast traffic. The configuration includes the configuration for RRC_INACTIVE state as well as the PTM Group-RNTI and RMA Id consisting of at least one cell. The anchor gNB receives the PTM multicast traffic over the N3 data tunnel from UPF. When the UE2 identifies a new cell with better coverage/quality and optionally the current source cell is having degrading coverage/quality, the UE needs to perform a cell reselection to a new cell. In Figure 5-2, two cases with UE mobility can be identified for UEs in RRC_INACTIVE state.
    a. The UE2 moves inside the RMA Id1 and performs cell reselection from one DU to another DU. The UE2 does not need notify the network about cell reselection since it is able to receive the same multicast traffic from all transmission points under the same RMA. The RMA can consist of one or more gNBs and the UPF traffic is distributed over F1, Xn and N3 interfaces to transmission points to cover the RMA area. Further, if the RMA consists of multiple gNB contributing to MC-PTM transmission, then the F1, Xn (necessary synchronization can be controlled by gNB receiving the IP multicast traffic over N3) and N3 interfaces can be used to route the same IP multicast traffic to joining gNBs.

b. If the new target cell is outside of the RMA Id1, UE needs to notify the network its new location with RMA update. Network will configure the UE with new RMA Id and if the new RMA consists of more than one gNB, network performs the RAN based Multicast Area Setup to allow traffic distribution over Xn to gNBs belonging to RMA.

3. The UE3 is having low unicast activity, its connection towards AMF is released and therefore the UE3 is configured with RRC_IDLE state. In this state the core network knows the UE's location only within the tracking area in AMF. Alternatively, the UE3 could be also a receive only device or in ROM mode without uplink capability, thus the network doesn't know its location. In these cases, the RMA may be configured with multiple cells participating in SFN broadcast mode. The RMA becomes the same as the tracking area or SFN service area and two or more selected cells are participating in SFN, for example according to given pre-configuration. When the area of RMA Id2 is configured with SFN transmission, all the UEs in that area can benefit from the SFN transmission regardless of their RRC state.

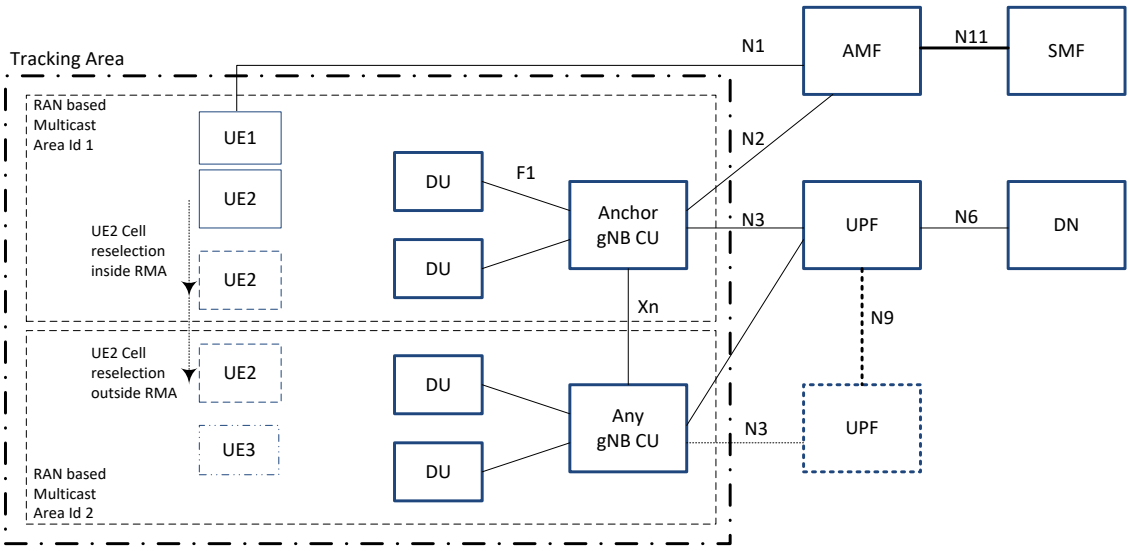


Figure 5-2: 5G RAN logical architecture and RAN based Multicast Area

Based on these RMA procedures, 5G-Xcast RAN architecture provides an option to configure the RAN for Terrestrial Broadcast distribution. In this case, users are unknown to the RAN (due to the lack of uplink and, therefore, registration into the network) and the RAN will decide beforehand and according to service and coverage requirements the multicast bearer configuration for delivery.

From the three examples shown above, Terrestrial Broadcast would be an extension of UE3 being it a receive only device with no uplink capability. In this case the RMA becomes the same as the tracking area or SFN service area. Single-cell transmission of a Single Frequency Network (SFN) with multiple cells participating can be configured.

5.2.2 RMA Setup

The RMA area may be dynamically managed over Xn interface between gNBs, when the RMA consists of cells belonging to more than one gNB. The UE can receive the multicast traffic and move within RMA without notifying the network, as long as it stays within the given RMA. When UE identifies a better cell out of the RMA, it will indicate the need for RMA update and performs cell reselection to a new cell. The requested RMA update can be UE specific or a multicast group specific or can consist of a list of multicast groups. This way, the multicast traffic delivery and RMA management can allow dynamic multicast areas, which are adapted to distribution of UEs in a geographical area.

5.2.3 RMA Update and Mobility

This subsection describes how to dynamically update RMA taking into account the challenges of UE measurement and mobility.

When a UE is in RRC_CONNECTED state, the UE can receive IP multicast data over default unicast Data Radio Bearer (DRB) or multicast DRB and mobility of the UE can be performed via conventional handover for UE in connected state. Regarding service continuity for RRC_CONNECTED UEs, the service interruption due to acquisition of target cell control information can be reduced by pro-actively sending the target cell control information to the UE via handover command similar to [44]. On the other hand, if low unicast activity is detected, the gNB can instruct the UE to suspend to RRC_INACTIVE state and the UE starts to be served by an updated RMA. Figure 5-3 demonstrates an example of RAN procedure for RMA updates with transition of one UE from connected mode to inactive mode. Herein, the gNB has detected that the UE1 has low unicast activity and the UE1 transitions to RRC_INACTIVE state and therefore the corresponding serving cell, cell 1, is added to the RMA configuration. As part of the RRC reconfiguration control, information such as Resume ID, Group-RNTI and the RMA updates, are sent to the UE. Accordingly, transmission of IP multicast data to the UE is performed over the updated RMA.

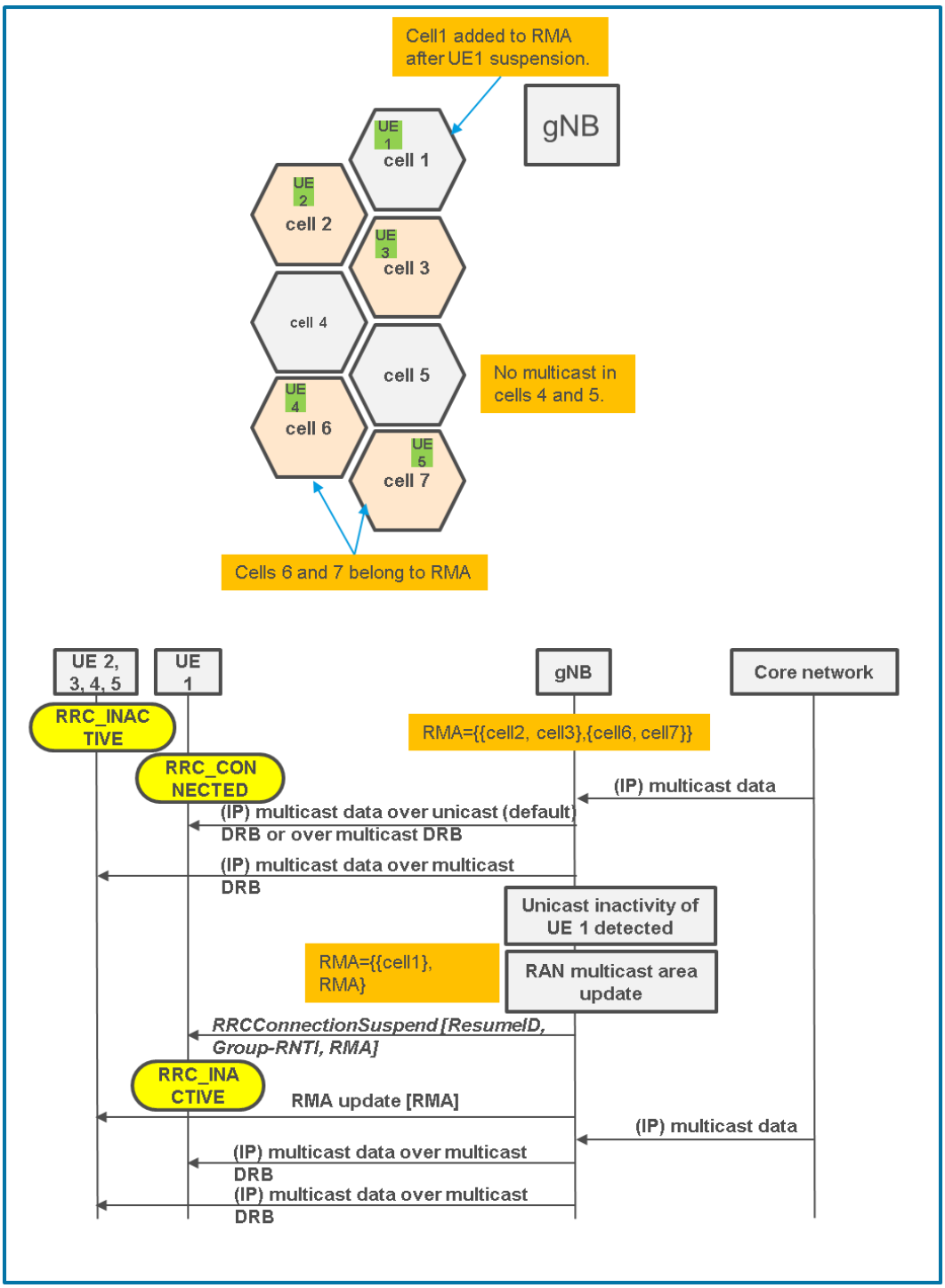


Figure 5-3: An example of RAN procedure for RMA updates with transition of a UE from connected state to inactive state.

Regarding mobility of UEs in RRC_INACTIVE state, UE-based or gNB-assisted update of RMA can be performed. In the case of gNB-assisted update of RMA, the UE should provide (via notification messages) reference signal measurements (RSRP/RSRQ), used for cell selection/re-selection, to the gNB to help RMA update decisions. The notification messages can be similar to RAN-based notification area updates, as specified in [11] or via measurement report (commonly used after transition to

RRC_CONNECTED state). Detailed procedures in RRC_INACTIVE state to perform the notification message that includes measurements are for further study. In particular, use of measurement report in RRC IDLE and RRC_INACTIVE states is not supported in the current 3GPP specifications [11]. In the following description, relevant UE measurement reports are notified to gNB after UE's transition to connected state.

Typical challenge with UE measurement is that, instantaneous UE measurements suffer from fluctuations due to fast fading and noise. The UE measurement fluctuations could lead to unstable RMA update decisions by the gNB. Hence, 3GPP has specified infinite impulse response filtering [16] to smoothen such fluctuations. However, filtering of UE measurements introduces delays with respect to the ideal channel we want to measure. Figure 5-4 demonstrates an example of instantaneous UE reference power measurement (pink curve), filtered measurement (green curve) and ideal measurement (blue curve) we want to achieve [46]. The filtered measurement uses 3GPP-compliant L3 filtering with time constant of 100 ms. Herein, it is observed that the delay of the filtered measurement can lead to delayed RMA update decisions by the gNB. As a result, pro-active preparation of target cell for RMA update is indispensable.

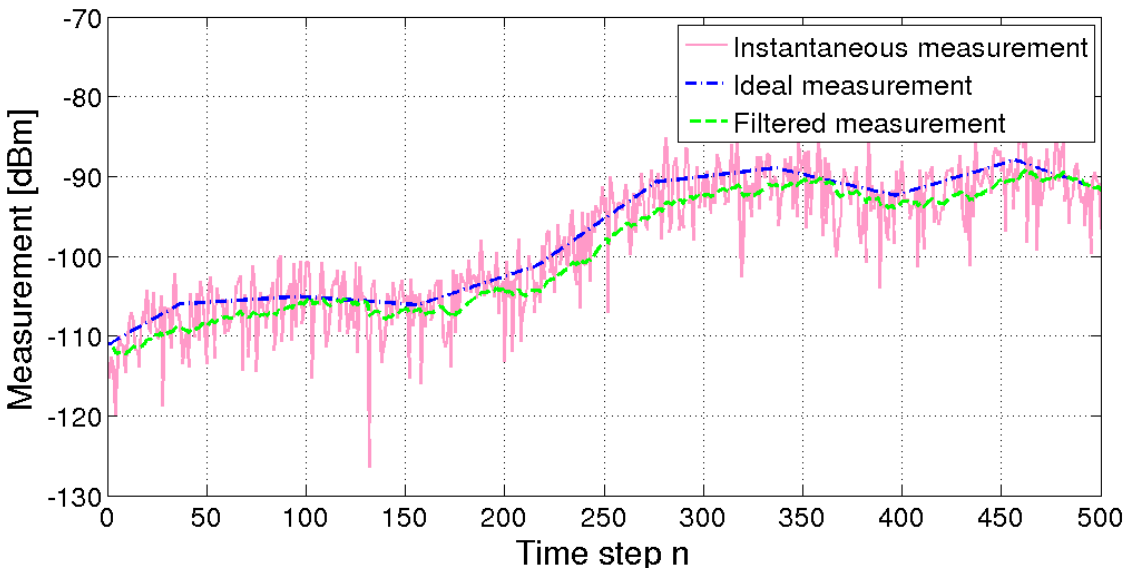


Figure 5-4: An example of instantaneous UE reference power measurement (pink curve), filtered measurement (green curve) and ideal measurement (blue curve) we want to achieve [46].

One of the criteria for pro-actively adding a cell to RMA is the comparison of the filtered UE measurement of the target cell against the difference between filtered UE measurements of the cells in the current RMA and certain measurement offset. The measurement offset parameters as well as the trigger mechanism (for the UE to generate and send measurement report to the gNB) can be configured by the gNB. Figure 5-5 describes an example of a procedure that performs RMA updates for mobility of a UE at RMA borders. Herein, sample UE 2, which has mobility at RMA border towards cell 4, send measurement report (after transition to RRC_CONNECTED state via connection resume request/response procedures) when the filtered measurement of the target cell is within a certain window with respect to the strongest filtered measurement in the current RMA cells; i.e., if filtered UE measurement of target cell is above the difference between the strongest filtered measurement and certain measurement offset for a certain time to trigger. From the measurement report, the gNB detects the strong neighbour cell, cell 4, and updates the

RMA by adding cell 4. If there is low unicast activity at the UE, the gNB send RRC suspend message (with the RMA re-configurations) to UE 2. Besides, the gNB sends reconfiguration message to the other UEs being served by the RMA, UEs 3, 4, and 5, to provide information about the new RMA update. Accordingly, the UEs update their RMA re-configuration information and starts to receive IP multicast data over multicast DRBs.

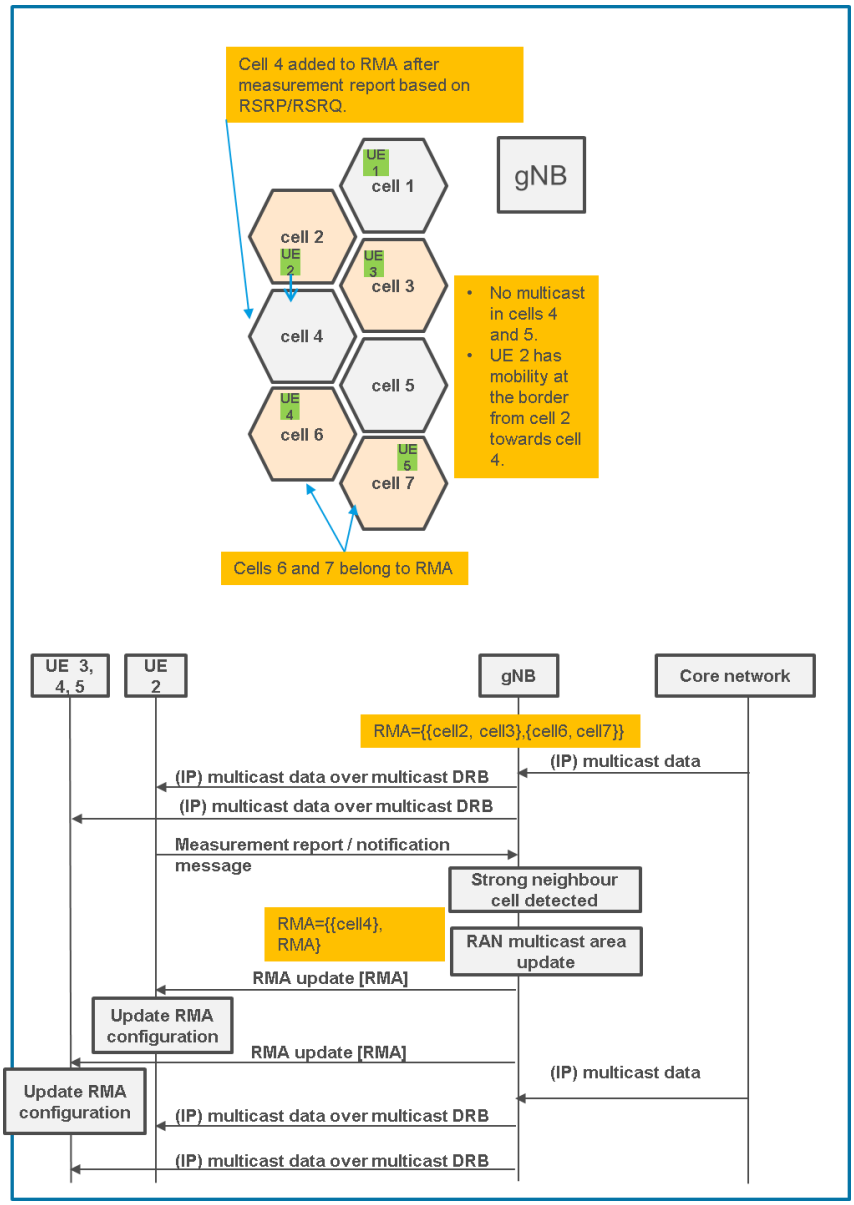


Figure 5-5: An example of a procedure that performs RMA updates for mobility of a UE at RMA borders.

It can be the case that some member cells in the RMA are serving one or more UEs with weak signal, in particular with mobility of a UE away from the cell or due to large scale fading processes. Such cells should have a mechanism to be removed from the RMA. The trigger criteria for sending notification message or measurement report can be checking if the difference between the filtered UE measurement of the strongest and weakest cell in the RMA is higher than a certain threshold. Similar to previous procedure, the threshold and trigger criteria can be configured by the gNB when

measurement controls are sent to the UE. Figure 5-6 demonstrates an example of RAN procedure to remove a cell for a UE. In this case, UE 2 sends notification message or measurement report (after transition to RRC_CONNECTED state) when the difference between the weakest and strongest filtered UE measurement in the RMA is higher than certain threshold. If there is low unicast activity at UE 2, the UE state is reverted back to RRC_INACTIVE. If UE2 is the only UE being served by the RMA, the gNB removes the weakest cell and updates the RMA. Accordingly, reconfiguration message is sent to the UEs being served by the RMA, UEs 2, 3, 4, and 5, to provide information about the new RMA update. Consequently, the UEs update their RMA re-configuration information and continues to receive IP multicast data over multicast DRBs.

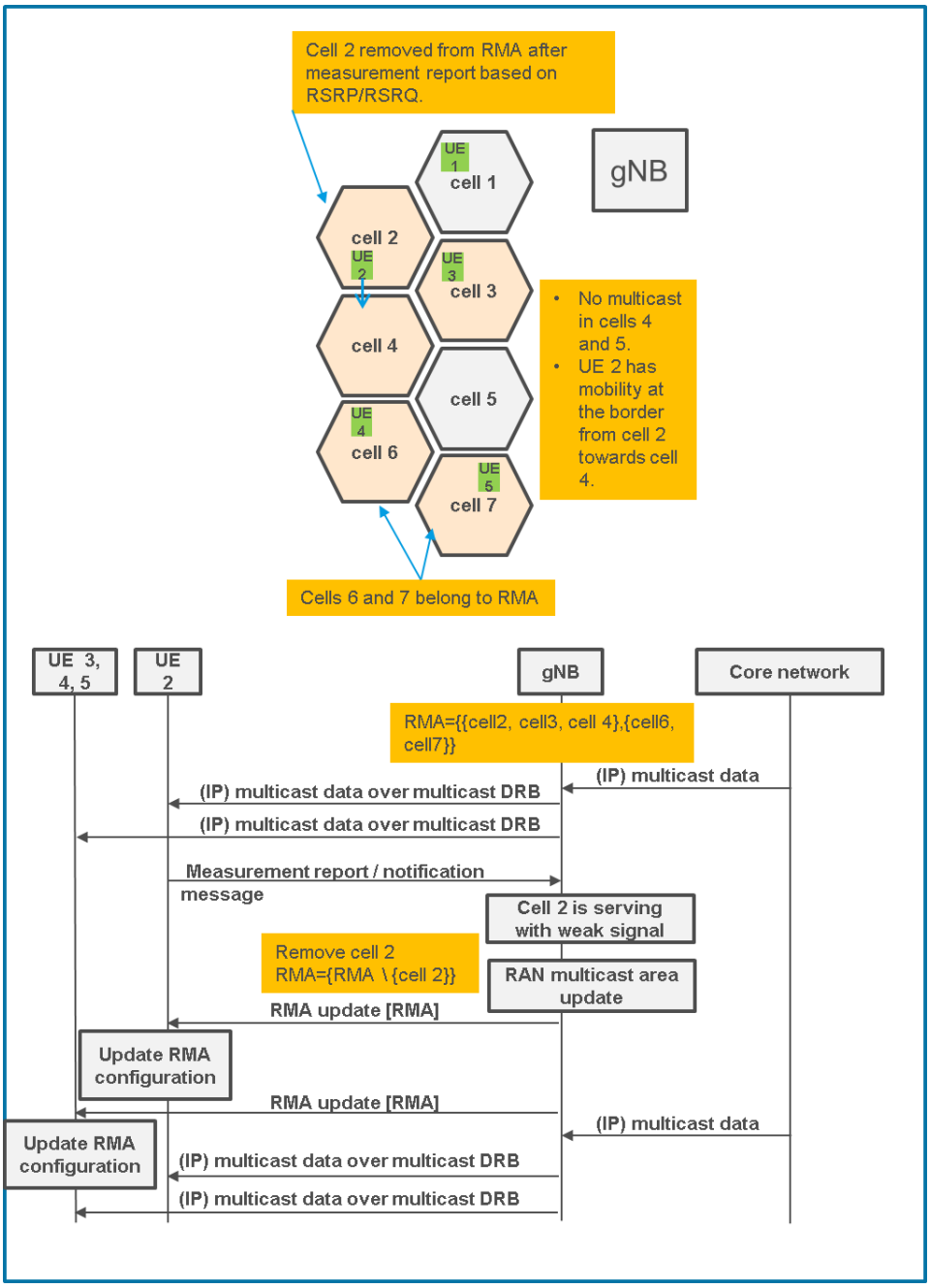


Figure 5-6 An example of RAN procedure to remove a cell from RMA.

5.2.4 SON functions for RMA

Self-Organizing Network (SON) functions have been introduced in LTE to reduce the operating expenditure (OPEX) required to manually configure large and complex networks as described in 3GPP TS 36.300 [9]. These functions are mainly categorized as self-configuration, self-optimization and self-healing functions. Self-configuration functions are defined as the processes where newly deployed nodes are configured by automatic installation procedures to get the necessary basic configuration for system operation. On the other hand, self-optimization functions are defined as the processes where UE and eNB measurements and performance measurements such as Key Performance Indicators (KPIs) are used to auto-tune the network. Besides, self-healing functions encompass mechanisms for automatic detection and localization of failures and techniques for healing the detected failures. Further technical details of SON function in LTE can be found in [61][62][63][64][65]. In addition to the classical SON functions (such as automatic neighbor relations, mobility robustness optimization, mobility load balancing, capacity and coverage optimization, inter-cell interference co-ordination, etc.), SON has shown to be a prospect in the design component of enhanced Coordinated Multi-Point (eCoMP) [66] and Active Antenna Systems (AAS) [67][68].

SON functions are expected to be crucial component of 3GPP's 5G networks, as demonstrated in SP-180827 [69]. One of the main reasons is that 5G networks will have many more cells and will be more complex than previous generations of cellular networks, and SON automation is even more important. According to SP-180827 [69], 5G SON is expected to utilize network management data, including, but not limited to, alarms, measurements, analytical KPIs, QoE, and provisioning data to analyze the network behavior, status, and traffic pattern, based on time and locations, to predict the potential issues, and to plan a solution in advance to resolve the issues before happening.

Dynamic and flexible RMA is crucial component of 5G-Xcast RAN design and the major drivers for RMA SON functions include, but not limited to,

- Temporal and spatial change of concentration of new traffic or UEs which may call for dynamic switching of PTP and PTM transmission to serve low and high traffic volumes, respectively.
- Propagation conditions, for instance, a change in the environment due to construction of new buildings or streets, dense leafy trees in spring season and falling leaves in autumn.
- Deployment of new radio access node which changes the overall network structure and this calls for adjustment of network parameters accordingly.

Once a radio access network is rolled out with the optimal configuration of RMA parameters such as thresholds for updating (adding, removing or modifying) RMA cell list, change in environment or traffic concentration calls for adjustment of the RMA parameters autonomously. The auto-tuning or optimization of the RMA parameters can practically be realized by using SON functions.

In accordance with centralized, distributed and hybrid SON architectures [61], the potential implementations of SON functions for RMA are demonstrated by higher-level figures below. Figure 5-7 and Figure 5-8 demonstrate centralized SON architecture for CU-DU deployment and distributed RAN, respectively. Herein, the SON algorithms for RMA, which includes processing of big data of UE and gNB measurements as well RMA parameter optimization, can be performed at the Operation, Administration and

Management (OAM) in the network management level. Centralized SON helps in accumulating comprehensive information from the network and it allows combining a wide range of information available at OAM to provide optimal RMA parameters. However, the major drawbacks of centralized SON architecture include longer response time of the RMA parameter optimization algorithms and network overhead on the communication between the gNBs and OAM.

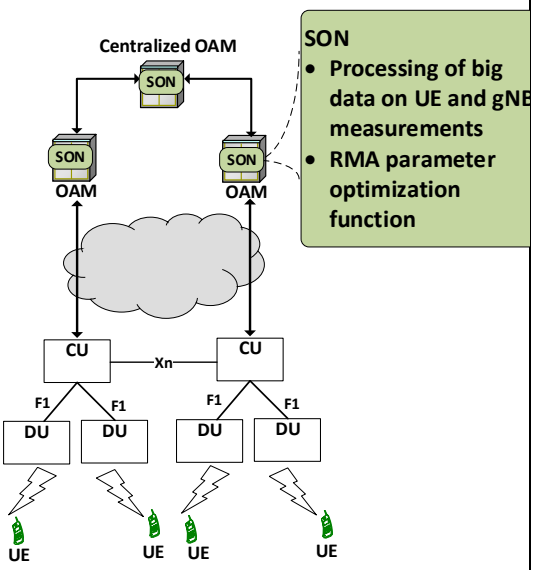


Figure 5-7: Centralized SON for RMA in CU-DU deployment

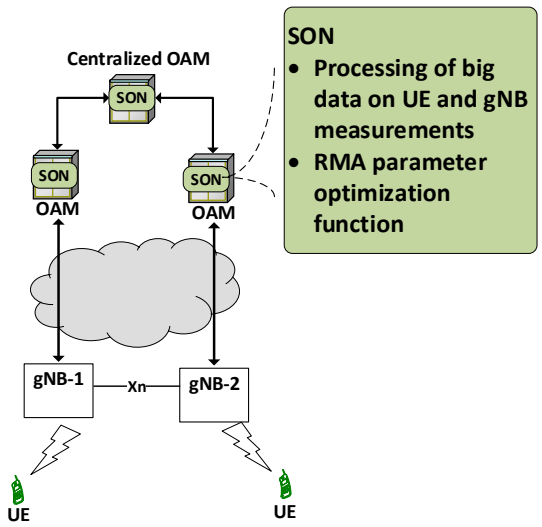


Figure 5-8: Centralized SON for RMA in distributed RAN

Figure 5-9 and Figure 5-10 show distributed SON architecture for CU-DU deployment and distributed RAN, respectively. With the distributed SON architecture, the SON algorithms for RMA parameter optimization are performed at the CU in case of CU-DU deployment and at the gNBs in case of distributed RAN. In the CU-DU deployment, installing the SON functions at the CU plays crucial role in coordinating intra-CU RMAs. For inter-CU RMAs, communication via Xn interface is needed. The major SON functions that can be considered in both CU-DU and distributed RAN deployments are processing of UE measurements (such as SS-RSRP, CSI-RSRP, etc.), processing of gNB measurements (such as count of UEs, KPIs for RMA optimization, etc.), the RMA parameter optimization function, and exchange of SON information (via Xn interface between gNBs in CU-DU deployment and between gNBs in distributed RAN). The major benefit of distributed SON architecture implementation is simplicity since it optimizes the RMA parameters locally without involving the OAM which in turn decreases the network overhead for the communication between gNBs and OAM. Moreover, the response time for RMA parameter optimization is fast. The drawback can be possibility of sub-optimal solution since all the information in the network is not available for the RMA parameter optimization.

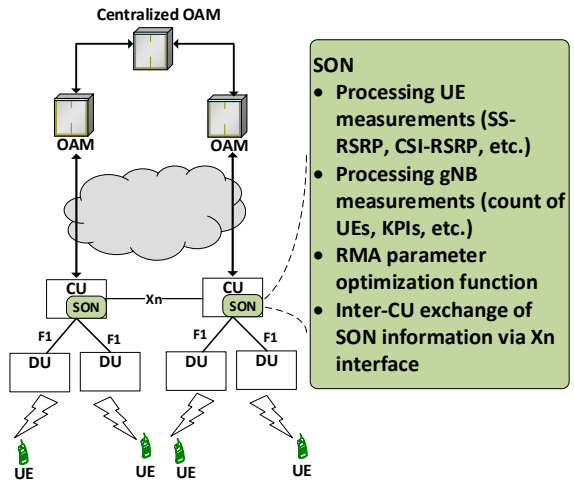


Figure 5-9: Distributed SON for RMA in CU-DU deployment

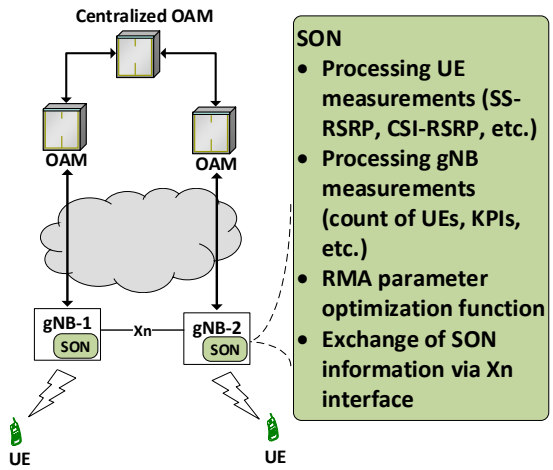


Figure 5-10: Distributed SON for RMA in distributed RAN

Figure 5-11 and Figure 5-12 show hybrid SON architecture for CU-DU deployment and distributed RAN, respectively, for SON implementation of RMA. The hybrid architecture may combine the benefits of central and distributed SON architectures. Regarding RMA parameter optimization, neighboring cells play critical role in RMA optimal parameter configuration; hence, the SON functions at the gNB may include processing UE measurements (SS-RSRP, CSI-RSRP, etc.), gNB measurements (count of UEs, KPIs, etc.), RMA parameter optimization function and Inter-CU or Inter-gNB exchange of SON information. On the other hand, the SON functions at the OAM can be used for performing computationally intensive functions (for e.g., if there are extremely big data of UE and gNB measurements that cannot be executed by the gNB). Moreover, the SON function at the OAM may include monitoring the Xn interface to make decision on whether the Xn link quality is stable for SON information exchange between distributed gNB or CUs. Furthermore, the OAM may also play the role of coordinating and controlling the SON functions at the gNB or CU, e.g., switching on/off the SON functions at the distributed gNBs or CUs. In comparison to the distributed SON architecture, the hybrid SON architecture still carries extra burden on the network for communication between OAM and gNBs or CUs.

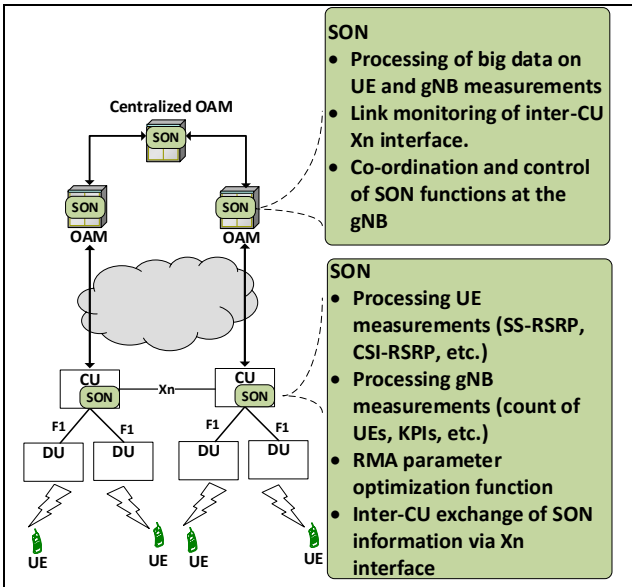


Figure 5-11: Hybrid SON for RMA in CU-DU deployment

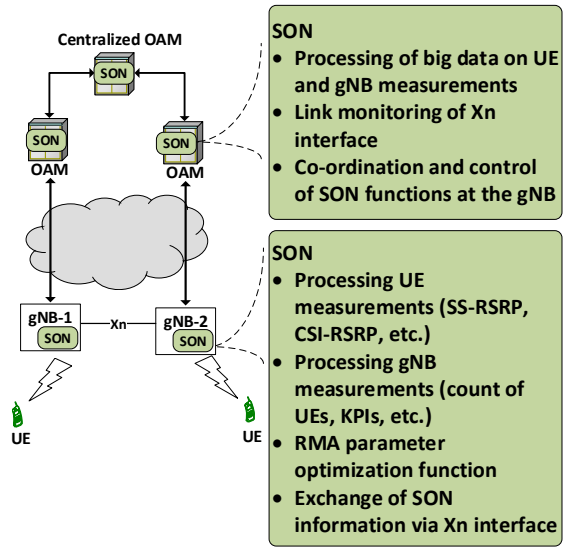


Figure 5-12: Hybrid SON for RMA in distributed RAN.

5.2.5 RAN Multicast Area Procedures for Terrestrial Broadcast

The RMA procedures for the setup of Terrestrial Broadcast transmissions leverages those already available for multicast in earlier sections. However, in this case, the RMA is configured according to pre-defined coverage requirements and agnostic to the QoS the UEs actually experiences (either they have uplink capabilities or not).

A Terrestrial Broadcast Service Area (TB-SA) is defined as the amount of time/frequency resources per transmitter area (either single transmitter or SFN area) reserved for the potential transmission of Terrestrial Broadcast services. In order to adapt to a variety of deployments suitable for the delivery of Terrestrial Broadcast services, the RAN is provided via O&M with the list of cells that constitute a given TB-SA with the following assumptions:

- Each single cell transmitter is considered as a constituent TB-SA
- A cluster of cells that constitute an SFN is regarded as a unique TB-SA
- A wide coverage area comprising a variety of topologies (e.g. mixture of single and SFN transmitter areas) is formed by means of multiple TB-SA.
- One transmitter can be operating more than a single carrier, therefore, each cell in the list may be associated a given frequency (DL_EARFCN).

Each TB-SA is identified by means of a TB-SA Index (TB-SA ID) which can be selected by the service provider via xMB [1]. XCF translates the RMA-TB indexes to the actual identifiers of the gNBs.

The amount of available resources per carrier might be different in each transmitter due to several circumstances such as the presence of other services, the use of carriers of different bandwidth or the needs of inter-transmitter scheduling (e.g. time/frequency reuse) to avoid interferences. Therefore, each TB-SA shall be informed of the specific amount of time/frequency resources that need to be available for potential service scheduling via xMB. It is a design assumption in 5G-Xcast architecture that services to be transmitted in SFN will be scheduled over dedicated resources with an adequate

numerology. The group of resources with different numerologies can be multiplexed by using different Carrier Bandwidth Parts (FDM-multiplexed within a given OFDM symbol) or subframes/frames (via TDM).

The delivery of each broadcast service, e.g. TV or radio, can be configured according to the TB-SA ID where the service is meant to be delivered. Associated to each broadcast service, the MCS index that fulfils the robustness (coverage) and data rate requirements of the SLA is indicated together with scheduling information in terms of required time/frequency resources for the given data rate (e.g. initial and final PRB).

An admission control procedure will determine the allocation of a new broadcast service according to the amount of available resources in the carrier for the allocation of Terrestrial Broadcast service (as indicated per TB SA) and the amount of required resources per service.

5.3 Selective FEC in 5G-Xcast RAN Architecture

5.3.1 Introduction

Forward Error Correction (FEC) framework is used at various layers of protocol stack. Unidirectional communication often uses FEC at higher layers to minimize impacts of packet loss between the communication end-points. One example of system using FEC at higher layers is enhanced Multimedia Broadcast and Multicast System (eMBMS). eMBMS utilizes File Delivery over Unidirectional Transport (FLUTE) for various MBMS user services such as 3GP-DASH, file download 3GPP TS 26.346 [70]. Internet Engineering Task Force (IETF) document RFC 6726 [71] builds on top of asynchronous layer coding (ALC) of document RFC 5775 [72] which combines layered coding transport, congestion control and forward error correction building blocks.

IETF describes a more general FEC framework for applications over public and private IP networks in RFC 6363 [73]. The FEC framework supports applying FEC to arbitrary packet flows over unreliable transport and is primarily intended for real-time or streaming media.

RTP allows for retransmissions as a repair method for streaming media. Multiplexing of original and retransmission streams is achieved using session-multiplexing for multicast and SSRC-multiplexing for unicast sessions. In session-multiplexing, RTP sessions for original and retransmission streams are sent on different network addresses/ports in document RFC 4588 [74].

In the FEC framework described in RFC 6363, the source and repair data can be multiplexed using RTP multiplexing in document RFC 4588. The FEC framework outputs are FEC source and repair packets. The FEC framework configuration information includes the definition of flows (e.g. port(s) and multicast address group(s)) for the FEC source and repair packets. The source flow can be the same as Application Data Unit (ADU) flow (e.g. User Datagram Protocol (UDP) source and target ports and source and target addresses of the IP datagram) as if the FEC framework was not present.

This section focuses on higher layer FEC in 5G architecture supporting multicast transport in which RAN functionality includes an autonomous determination to use unicast or multicast/broadcast transmission to efficiently deliver IP multicast data over unicast radio bearers. In Figure 6-3 5G-Xcast L2 architecture and bearer selection in Cloud-RAN, it is proposed that unicast and multicast/broadcast transmission can use different RLC entities driven by a switching function that can select between unicast and multicast transport channels.

Application of FEC to ADU flow increases the amount of user payload that the 5G system needs to transmit. FEC protects the ADU flow against packet loss, which is inevitable in unidirectional communication. In the context of 5G RAN architecture, RAN can decide whether to transmit multicast data using unicast or multicast/broadcast transmissions wherein said transmissions may have different configurations. Therefore, in the unicast case the lower layer HARQ or RLC retransmission may be used, which are more efficient compared to using higher layer FEC. Lower layer redundancy also requires less bits to be transmitted over the air compared to higher layer methods.

In the 5G-Xcast RAN architecture design, the RAN can select between unicast or multicast/broadcast transmission for multicast QoS flows, including means for the network to determine when such switch may be beneficial. However, RAN does not have information whether a QoS flow is used for FEC source or repair flow. The consequence is that the FEC repair flow will be transmitted over unicast when the lower layer retransmissions are more efficient.

5.3.2 Selective FEC architecture description

The overall selective FEC architecture is shown in Figure 5-13 where the application flow comprises of source and FEC flows, which are received by the UPF. In this example case it can be assumed that all the flows use the same multicast group and are differentiated by transport layer ports.

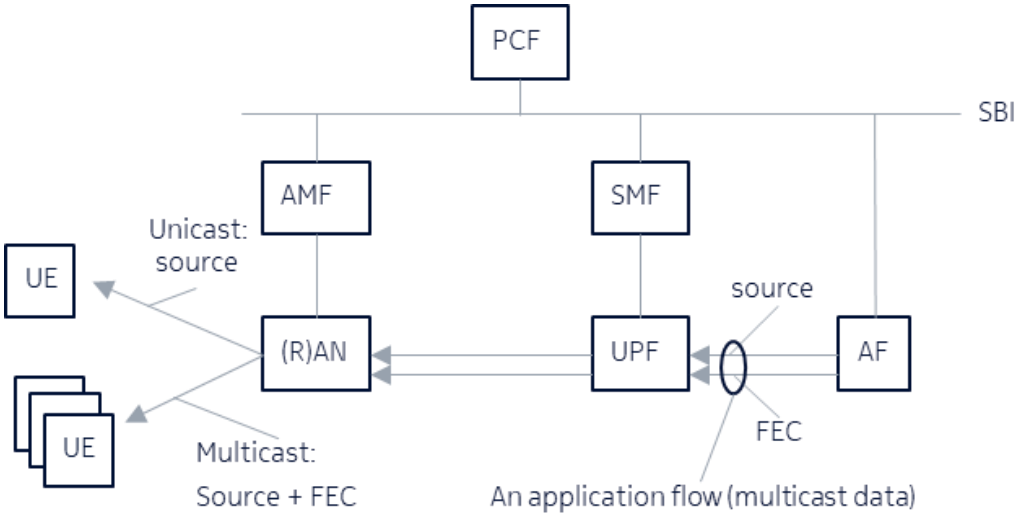


Figure 5-13: System architecture for application flow with unicast and multicast + FEC traffic.

In Figure 5-14, selective FEC is considered for both distributed RAN and centralized RAN. With distributed RAN, the gNB has the capability to decide the over-the-air transmission mode, and gNB-DU would be handling this role in the centralized RAN scenario transmitting FEC flows only in case multicast mode is used. For the unicast scenario, this saves significant amount of radio resources, while providing higher flexibility and performance to the RAN.

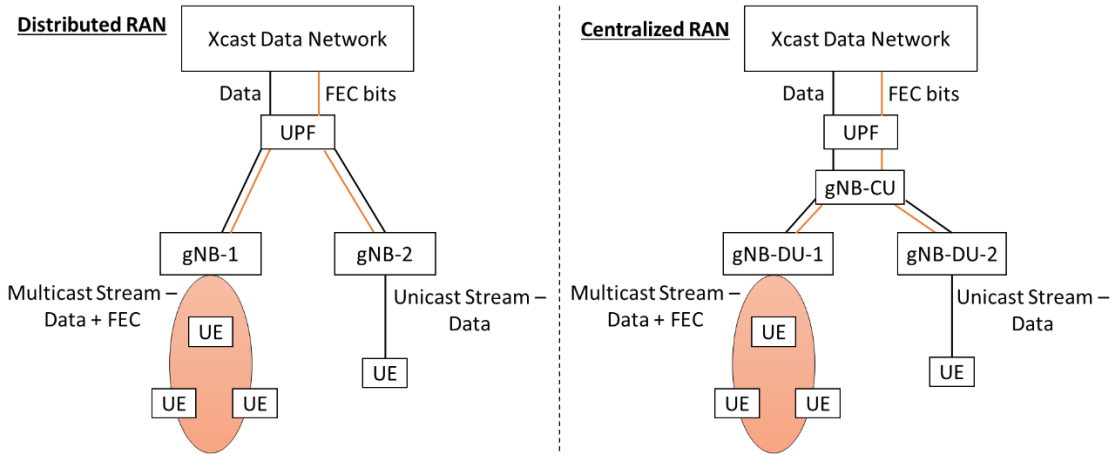


Figure 5-14: Operating the selective FEC in distributed and centralized RAN.

The protocol architecture is as shown in Figure 5-15, considering both distributed and centralized deployments with a mix of multicast and unicast flows. In this example the DU-1 of gNB-1 will schedule the source and FEC flows using Xcast radio bearers with multicast transmissions over-the-air. In another example the DU-2 of gNB-2 will be able to drop the FEC flow from transmission and schedule only the source flow using unicast transmissions over-the-air. These flows can be multiplexed with other existing unicast flows within the gNB / DU as well.

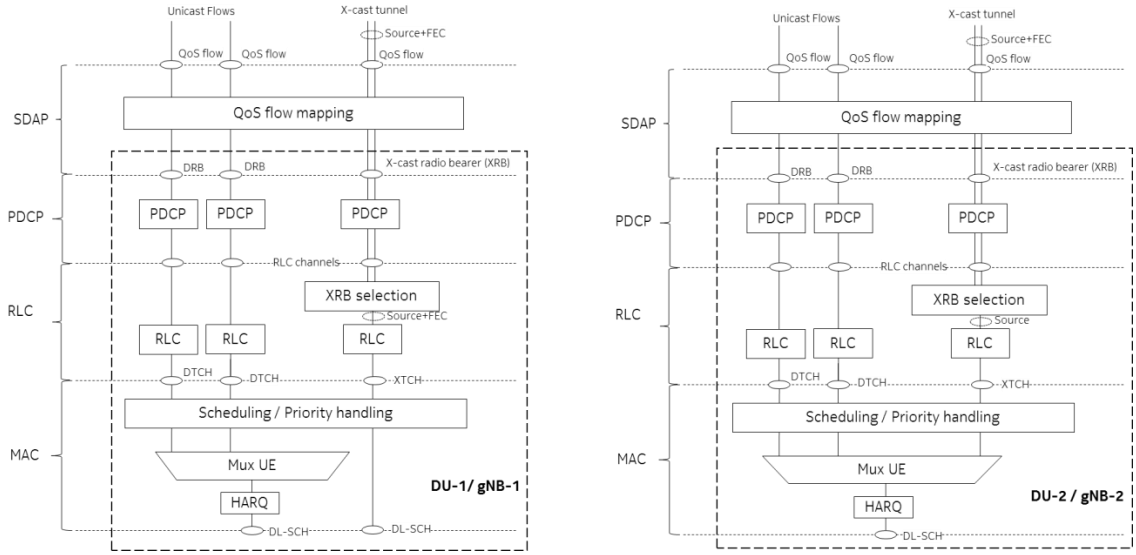


Figure 5-15: Selective FEC protocol architecture.

5.4 Summary

In this section we have introduced two novel architectural and functional enhancements to the 5G-Xcast concept. The RMA concept allows dynamic management of the multicast area along with UE mobility and connectivity. The dynamic areas are based on the user demand in receiving IP multicast traffic and taking into account the geographical distribution of users, reusing the flexibility of the unicast architecture and basic principles of SC-PTM extended over to Multi-Cell PTM (MC-PTM). It is proposed that the RAN can decide the multicast bearer configuration and deliver the multicast traffic over unicast data radio bearers when the number of users is low. The RMA,

where the UE can receive multicast traffic during the low activity unicast periods is defined and controlled anchor gNB (usually the last serving gNB). Depending on the number of low activity UEs receiving multicast in a cell, the gNB can decide to activate the multicast traffic delivery over PTM, MC-PTM or SFN transmission modes.

SON algorithms, such as self-configuration and self-optimization, were identified as suitable methods in configuring and maintaining the dynamic and flexible operation of RMA. It was found out that the temporal and spatial change of new multicast traffic and/or UEs calls for dynamic switching of PTP, PTM and MC-PTM transmission to serve low and high traffic volumes. Further, the RMA may need to adapt to the deployment of new radio access nodes or propagation environment changes and therefore the adjustment of transmission mode parameters using SON methods seems a practical way of managing the 5G-Xcast networks.

Finally, the application of FEC to protect the application traffic flow against packet loss in air interface in unidirectional communication was studied. In the context of 5G-Xcast RAN architecture, the transmission mode selection was introduced to allow the RAN to use the most efficient transmission method allowing saving of radio resources. When the RAN decides to transmit the multicast data using unicast PTP, the lower layer HARQ or RLC retransmission may be used, which is more efficient compared to higher layer FEC.

6 5G-Xcast RAN Deployment and Functions

6.1 5G-Xcast deployment scenarios

The 5G-Xcast RAN deployment leverages the major assumptions of 5G-NR overall architecture described in [17] which shows RAN architecture for gNBs with and without functional splits.

6.1.1 RAN deployment without functional split

With this deployment scenario, all the logical gNB functions as well as RAN interface protocol terminations are hosted in a gNB physical node. Figure 6-1 depicts 5G-Xcast RAN deployment scenario without functional split. Herein, the logical nodes include CP and UP. The UP hosts 5G-Xcast control functions including functions performed by gNB-CU-MC (section 3.2.2). On the other hand, the UP logical node hosts 5G-Xcast RAN function for delivery of user plane data. The major interface protocol terminations for 5G-Xcast interfaces are NG-C (to which N2 reference point is mapped), NG-U (to which N3 reference point is mapped), M1-NG, Xn-C and Xn-U. Details of the interface protocols are described in Section 4. Accordingly, NG-C and NG-U are CP and UP of NG interface whereas Xn-C and Xn-U are CP and UP of Xn interface. M1-NG interface is optional interface and it is considered in the architecture Alternative 2 [1], where XUF is directly connected to the RAN.

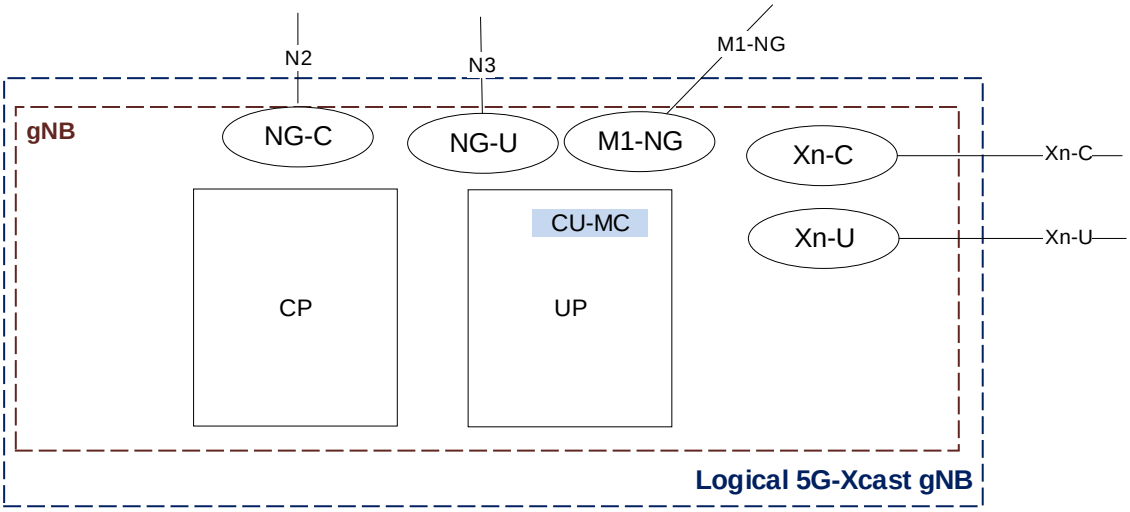


Figure 6-1 5G-Xcast RAN deployment scenario without functional split.

6.1.2 RAN deployment with functional split

In line with NG-RAN deployment scenario assumptions in [17], Figure 6-2 demonstrates 5G-Xcast RAN deployment scenario with functional split. Herein, the figure shows logical nodes (CU-CP, CU-UP and DU), internal to a logical gNB. The major interface protocol terminations for 5G-Xcast interfaces, NG-C, NG-U, M1-NG, Xn-C and Xn-U, are hosted in the central entity. The DU is hosted in a distributed entity. The central entity and distributed entity are separate physical nodes.

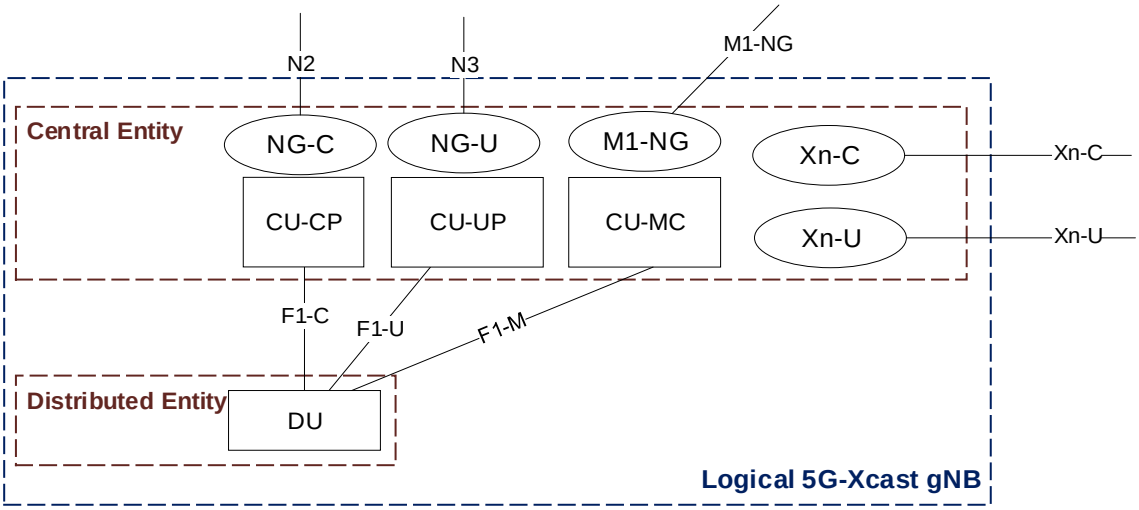


Figure 6-2. 5G-Xcast RAN deployment scenario with functional split.

6.1.3 Cloud-RAN based deployment

In a typical Cloud-RAN deployment, the functional split into CU and DU network entities allows RAN architecture with centralized processing of RAN functions with higher layer processing of SDAP and PDCP, e.g. according to Figure 6-2. At high level, the DU(s) closer to the deployed cells receive information about a set of UEs to which the multicast data should be transmitted and based on this information the distributed unit configures the needed unicast channels and multicast channels. The CU being a centralized unit and DU a local unit, the DU neds to make the decision of the transmission mode. When the DU receives multicast data from a CU, it will select either unicast or multicast channel to transmit the multicast data to the set of UEs as per the procedures described in Chapter 5.2, RAN Multicast Area procedures.

The proposed Layer 2 radio protocol architecture for cloud-RAN deployments is shown in Figure 6-3. The multicast data is delivered to NG-RAN over a data tunnel, which in this case is referred as X-cast tunnel in Figure 6-3 to emphasize the dynamic selection process of RLC entities and transport channels for the transmission. The multicast traffic can comprise of multiple QoS flows. In this case the SDAP can map the QoS flows to a set of X-cast data radio bearers (XRB) to enable differentiation at lower layers for different QoS requirements.

The PDCP, which is not used in eMBMS architecture, may provide sequence numbering and duplication detection. In case the UE is receiving the same data over unicast and multicast DRBs, the duplication detection should be supported. The duplication can be used in the proposed architecture also for performance enhancement when the UE receives the same PDCP PDU over DTCH and XTCH as a means for improving packet reliability. In this case the ciphering functionality used for unicast is not required for multicast. Even if this does not impose a security risk, it might be possible to discover which IP multicast group the UE is listening. Another PDCP function relevant to the transport of multicast data is the header compression and decompression.

Switching function in the DU is the new functionality proposed to the architecture, where the DU selects the transmission method. Switching function locates below PDCP but above RLC layer, thus not placed in the same Cloud-RAN computing

hardware pool as the CU. Thus, using the F1 fronthaul interface, it is natural to place the Switching function in the DU. For a set of UE's receiving multicast data (i.e. the UE's which have expressed their interest in receiving multicast and the PDU session modification procedure has been completed), a pair of RLC entities and logical channels (i.e., DTCH and XTCH channels) is set up to transmit the multicast data over the air. The multicast logical channels are shared between some or all multicast UEs.

The switching between unicast and multicast can be based on the availability of UE measurements and the reported quantities in the measurements, such as SS-RSSP, CSI-RSRP, SS-RSRQ, CSI-RSRQ, according to procedures related to RAN Multicast Area in Chapter 5.2. In general, if measurements are not available, XRB switching is routing traffic though multicast transport channels and when measurement reports indicate poor radio condition for some UEs in comparison to others, the Switching function will select the unicast transport channel for those UEs and the multicast transport channel for other UEs. When setting up an XRB, RRC may configure thresholds in the XRB switching function to select between unicast and multicast logical channels, also considering the minimum number of UEs required for switching to multicast transport and the resulting estimated resource and spectrum efficiency gain.

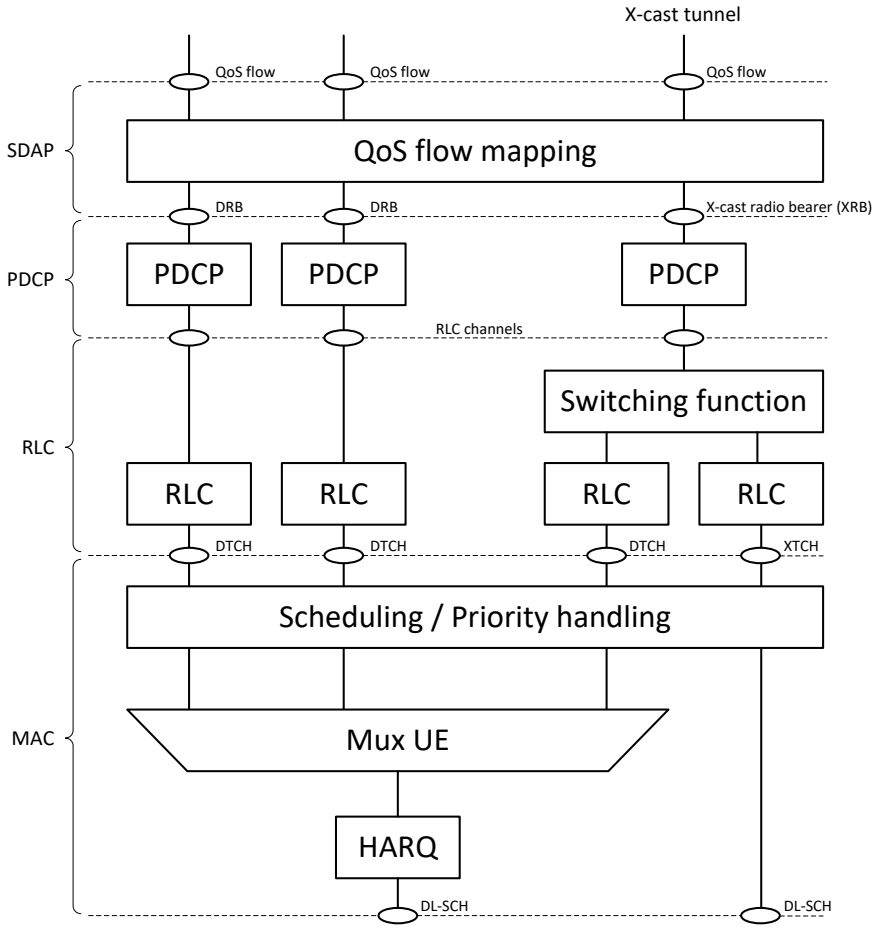


Figure 6-3 5G-Xcast L2 architecture and bearer selection in Cloud-RAN

6.1.4 PDU Session Modification Procedure

The PDU Session Modification procedure is described in this section. The purpose of the PDU session modification is to allocate the RAN resources for UEs joining in the IP multicast group. The Cloud-RAN based L2 architecture was shown in Figure 6-3 and the following call flow in Figure 6-4 is assuming the same architecture. The PDU Session Modification procedure is described for the first UE joining the IP multicast group. The same procedure is performed for all UEs, even if some information may be optional for some subsequent UEs joining the IP multicast group.

1. PDU SESSION RESOURCE MODIFIY REQUEST from AMF to gNB-CU (TS 38.413): The message is modified to include multicast context related configuration including: UP transport layer information for multicast traffic, multicast address identifying multicast group UE wants to receive traffic from or an alias ID (e.g. multicast context ID), and QoS flows information (indicators, QoS parameters).
2. UE CONTEXT MODIFICATION REQUEST from gNB-CU to gNB-DU: When this call flow is performed for the first UE, gNB-CU allocates PDCP entity for new XRB with an identity (XRB Identity) and optionally an UL tunnel end-point for the XRB. gNB-CU then sends the UE CONTEXT MODIFICATION REQUEST message to gNB-DU including the UL tunnel end-point information (optional), XRB Identity, QoS information selected by gNB-CU based on the mapping of QoS flow to a data radio bearer at SDAP, and the switching configuration. The call flow for the subsequent UEs may include only XRB Identity.
3. The gNB-DU configures switching function, RLC channels and logical channels for XRB bearer and DL tunnel information. This includes the following steps:

    a. gNB-DU configures unicast transport by creating an RLC entity mapped to a single RLC channel towards PDCP and mapped to a corresponding logical channel in MAC according to DRB setup procedures.
    b. The switching function is configured according to the switching configuration, including thresholds for measurements and required minimum number of multicast UEs.
    c. If there is no multicast transport configured already (i.e. the same value of XRB Identity is not use part of any active UE context at the gNB-DU), a new RLC entity and corresponding mapping to a multicast logical channel (XTCH) is created. The RLC entity is mapped to the same RLC channel as in step 3a. The configuration includes at least one of the following: logical channel identities, RLC configuration (e.g. mode, sequence number field length, timer values), MAC configuration (e.g. XRB specific DRX as this may be different from UE's DRX) and PHY configuration (e.g. RNTI value for the reception of XTCH scheduled on DL-SCH).

    When gNB-DU receives a UE CONTEXT MODIFICATION REQUEST message which includes an XRB Identity already being associated with another UE, it indicates that there is a multicast transport configured already and the gNB-DU may not need to perform step c above but use the existing configuration.
4. UE CONTEXT MODIFICATION RESPONSE from gNB-DU to gNB-CU: gNB-DU reports the successful operation to gNB-CU and includes the configuration

information from step 3 in DU to CU RRC Information IE, including XTCH logical channel identity.

5. DL RRC MESSAGE TRANSFER from gNB-CU to gNB-DU: RRC takes into account the response from gNB-DU on logical channel configuration. RRC Connection Reconfiguration includes information on logical channel mapping by adding the configuration of multicast transport in the RLC-Bearer-Config IE: logicalChannelIdenity is set to the XTCH channel identity, and servedRadioBearer is set to XRB identity configured in step 3. For the configuration of unicast transport in the RLC-Bearer-Config IE: logicalChannelIdentity is set to DTCH channel identity, and servedRadioBearer is set to the same value of XRB identity.
6. RRC Connection Reconfiguration: UE performs DRB Reconfiguration as described in [16]. Two RLC-Bearer-Config IEs are present with the same servedRadioBearer identity, and if an XRB Identity is present, UE configures this bearer to deliver PDUs from both logical channels in RLC layer to one PDCP entity. This procedure is transparent to PDCP.

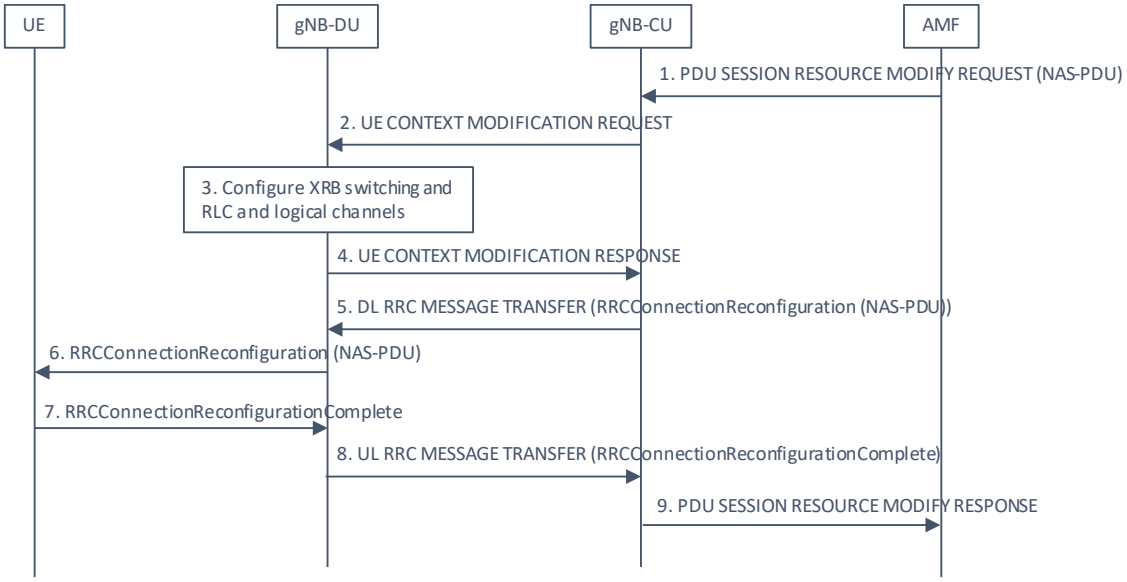


Figure 6-4 5G-Xcast PDU Session Modification for multicast resource allocation

6.1.5 Terrestrial Broadcast deployment

Some examples are provided regarding the cell arrangements from which a broadcast service may be transmitted, in this case assuming TV/radio services. In the following Figure 6-5, three different deployments are shown consisting of a nation-wide SFN, a regional SFN and a deployment covering the same area by means of single cell transmitters. A central hexagon is highlighted, which belongs to different TB-SA IDs according to the network planning requirements of each TV/radio service.

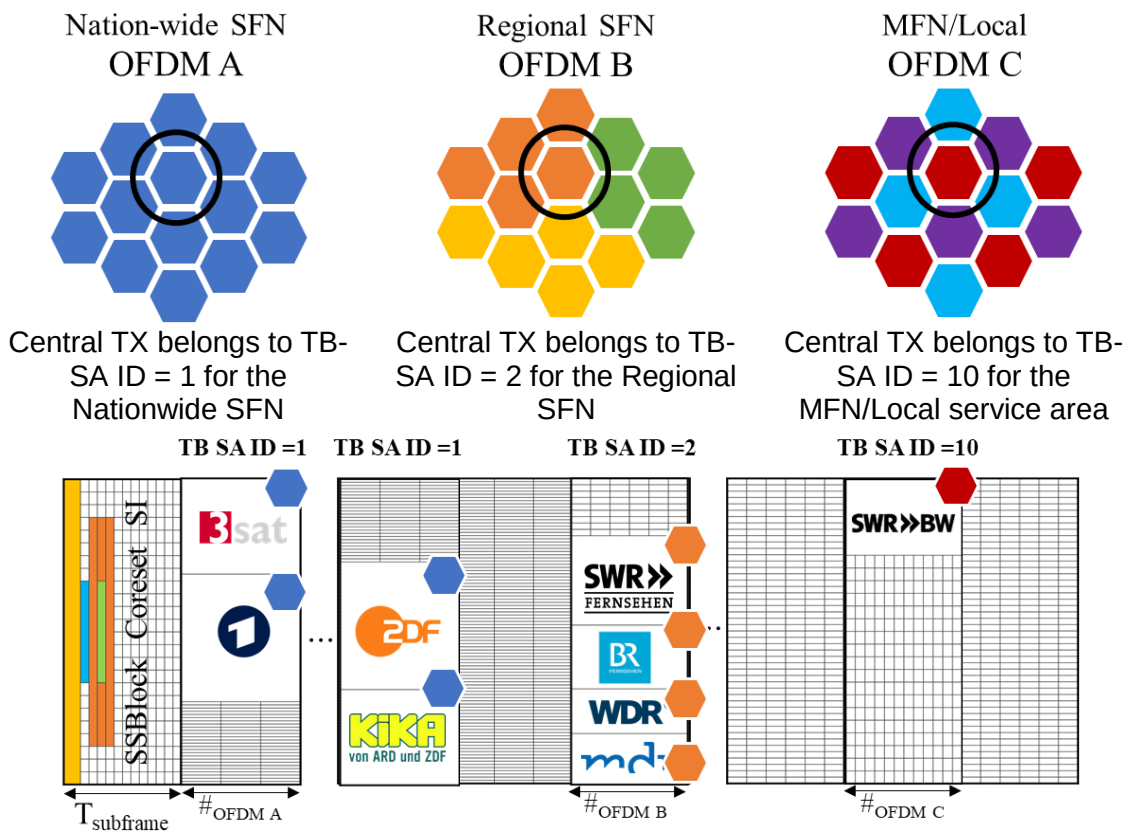


Figure 6-5 Three deployments consisting of a nation-wide SFN, a regional SFN and a single cell transmitter and their association to Terrestrial Broadcast Service Areas.

A frame transmitted from the central hexagon is shown, where, for simplicity, TDM is used to multiplex frames containing the services per different TB SA. The three scenarios are:

- A set of transmitters configured within the same SFN area. In this case a complete carrier (or frame within a carrier) is available to schedule Terrestrial Broadcast services
- A set of transmitters that constitute different SFN areas requiring synchronization and orthogonal scheduling between SFN areas
- A set of single-cell transmitters requiring orthogonal scheduling of resources to avoid mutual interference (in this case on a reuse 3 basis).

6.2 Logical placement of Network Functions

5G-Xcast RAN functions include 5G-NR functions defined for PTP in [11] as well as 5G-Xcast multicast and broadcast supporting functions. In case of RAN deployment without functional split, as described in Figure 6-2, all the RAN functions are placed in one physical gNB. On the other hand, for RAN deployment in Figure 6-2 with functional split, the logical placement of RAN functions at the CU and DU entities is depicted below.

Functions at CU

- Functions for Radio Resource Management: Radio Bearer Control, Radio Admission Control, Connection Mobility Control;

- IP header compression, encryption and integrity protection of data;
- Selection of an AMF at UE attachment when no routing to an AMF can be determined from the information provided by the UE;
- Routing of User Plane data towards UPF(s);
- Routing of Control Plane information towards AMF;
- Connection setup and release;
- Scheduling and transmission of paging messages;
- Scheduling and transmission of system broadcast information;
- Measurement and measurement reporting configuration for mobility and scheduling;
- Transport level packet marking in the uplink;
- Session Management;
- Support of Network Slicing;
- QoS Flow management and mapping to data radio bearers;
- Support of UEs in RRC_INACTIVE state;
- Distribution function for NAS messages;
- Radio access network sharing;
- Dual Connectivity;
- Tight interworking between NR and E-UTRA.
- Multi-cell synchronization;
- Dynamic RMA;
- RAN-level flexible transition between PTP and PTM radio bearers;
- Reception of MBMS control plane data at the NG-C interface termination;
- Reception of MBMS user plane data at the NG-U interface termination for 5G-Xcast mobile core architecture alternative 1 and alternative 3;
- Reception of MBMS data information at M1-NG termination point for 5G-Xcast mobile core architecture alternative 3;
- Radio resource management with flexible multiplexing of PTP and PTM radio bearers via dynamic scheduling in time, frequency and space;

Functions at the DU

- Functions for Radio Resource Management: Dynamic allocation of resources to UEs in both uplink and downlink (scheduling); Segmentation and multiplexing of protocol data units;
- Reception of control plane data at the F1-C interface termination;
- Reception of user plane data at the F1-U interface termination;
- Functions for 2nd layer of FEC with support for re-transmission of FEC protocol data units;

6.3 Summary

In this section the options of deploying 5G-Xcast RAN using the principles of NG-RAN were studied. It was shown that the 5G-Xcast RAN architecture can be deployed with distributed gNBs (e.g. 3GPP Release 15 with extension from unicast to multicast without split) and with CU/DU functional split for both IP multicast and terrestrial broadcast. The NG-RAN split option into CU/DU architecture allows more efficient placement and centralization of higher layer computing resources in CUs and at the same keeping time flexibility in choosing the transmission mode as unicast, multicast or broadcast at DUs depending on the requirements and factors discussed as part of RAN architecture requirements and RAN logical architecture in Section 2 and 3 respectively.

7 5G-Xcast RAN Slicing

7.1 Introduction

Network slicing is introduced to provide additional means to ease the telecommunication service provisioning and management in 5G networks. Network slicing is not so much technology driven as primarily business driven [47]. The key drivers are, to name a few, the need for the rapid deployment of new services, the need for the support for different operational models, and the need for accommodating conflicting functional requirements, etc. [48]. Network slicing is an important feature of 5G systems. 3GPP has been working on network slicing since Release 14 and giving definitions in different Technical Reports (TRs) and Technical Specifications (TSs) [7][49][50][51][52][53][54]. The latest 3GPP definition of network slice in Release 16 is "a set of network functions and corresponding resources necessary to provide the required telecommunication services and network capabilities" [55]. Here in 5G-Xcast we describe network slicing further with reference to [56][57]: a Network Slice is a virtual network created by the network operator, customized to provide an optimized end to end solution for a specific market scenario which sets specific requirements involving Core Network (CN) and Access Network (AN), where each slice supports one or more communication services. Network slicing can also be seen as an extension to the comparatively static principle of network sharing as specified in [58][59]. In light of that, RAN is an indispensable part of a network slice defined for PLMN, and in 5G systems.

In 5G-Xcast we follow the concept specified in 3GPP and provide end to end solutions for network slicing to support broadcast and multicast in 5G systems. The core network part of network slicing is discussed in [48]. In this chapter, we discuss the RAN part, specifically for 5G-Xcast traffic differentiation, traffic throttling and the application of RAN slicing to the RAN Multicast Area (RMA). 5G-Xcast RMA network slicing is the exact solution to meet the requirement specified in 3GPP on 5G Multimedia Broadcast/Multicast Services (MBMS), to "support Multicast/Broadcast network sharing between multiple participating MNOs, including the case of a dedicated MBMS network" [45].

7.2 Differentiated scheduling of 5G-Xcast traffic using RAN slicing

As described in the Sec. 3.2.2, the scheduling of multicast / broadcast traffic is done statically based on the synchronized pre-configurations within the 4G RAN and core network. Such static configurations might be applicable for wide-area terrestrial networks with pre-configured transmission areas – similar to terrestrial broadcast networks. In 5G, especially in the context of mixed-multicast mode, the key design principle adopted is to enable the RAN to decide how to schedule multicast/broadcast traffic, which could be done locally within a cell or a configurable area, which could be within the same gNB (assuming a gNB-CU with multiple DUs). The over-the-air transmission type (unicast, multicast, broadcast bearers) could be determined based on various factors such as user density, distribution, mobility, radio link conditions, etc. This means that the traditional scheduling configurations for multicast / broadcast traffic using static radio parameter and timing configurations cannot be applied in 5G.

Network congestion adds an additional dimension of complexity, since with the straightforward solution, the network would be throttling the traffic based solely on the traffic type similar to unicast traffic throttling applied today as shown in Figure 7-1. Traffic throttling based on QoS parameters (for e.g. 5QI priority level), which may also loosely indicate a traffic type (for example see example services of standardized 5QI

values - video, web browsing, file downloads, etc.), does not take into account the flexibility of dynamically configuring the transmission type for Xcast traffic, especially with different transmission types being used within the same gNB in different cells. Considering the fact that mixed-mode is mainly applied to over-the-top media traffic as compared to pre-configured broadcast media, the QoS-based solution provides sub-optimal performance for the network.

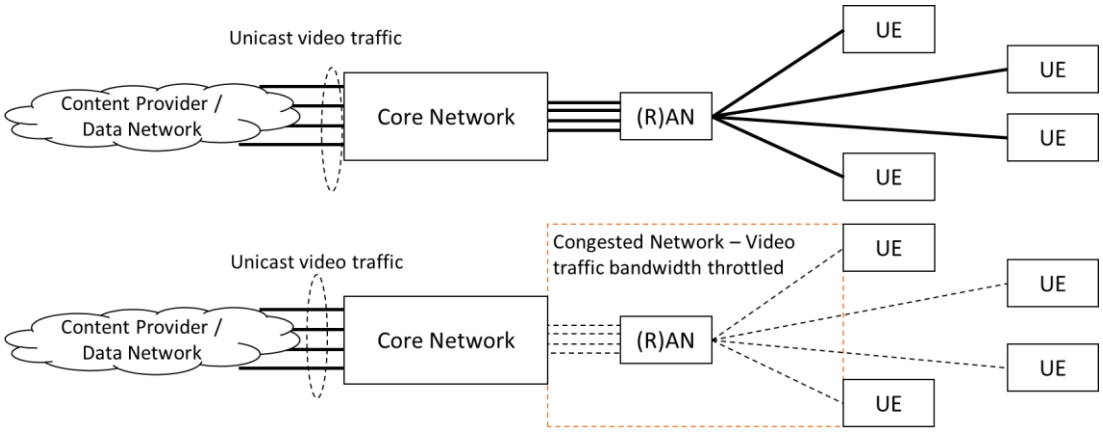


Figure 7-1 Traffic handling during congestion conditions

Consider the different scenarios shown in Figure 7-2, where Xcast and unicast traffic is simultaneously sent through gNB-{1-3}. The Xcast and unicast traffic represent different application session (e.g. consumption of two different live media streams). Xcast traffic is received by XC-UEs – with OTA transmission modes of unicast or multicast, and unicast traffic is received by UC-UEs. In 5G-gNB1, XC-UEs are scheduled using multicast with four UEs part of the multicast group. In 5G-gNB2, since there is only a single XC-UE, the Xcast traffic is scheduled using unicast, in which case no differentiated handling would be required. In 5G-gNB3, there are two XC-UEs receiving Xcast traffic using multicast transmission mode, with the UC-UEs in all the three gNBs always receiving the data using unicast transmission mode. Here, in the baseline scenario – without any congestion, the radio resources are scheduled according to the needs of the traffic volume that needs to be served for both the Xcast and unicast traffic.

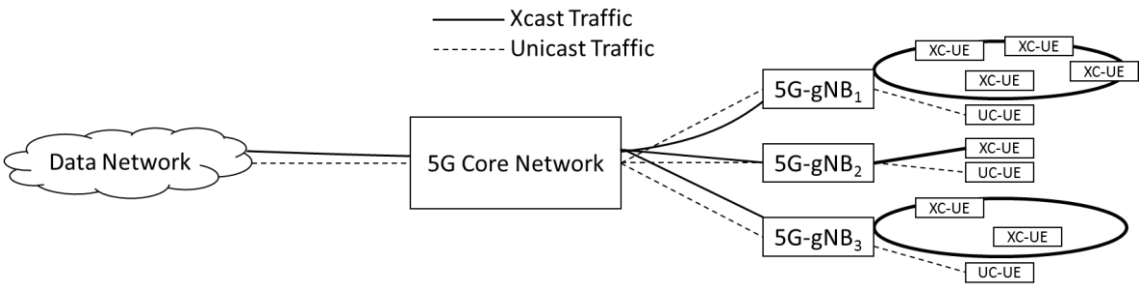


Figure 7-2 Different scenarios considered for differentiated scheduling

Here we discuss possible implementation of the differentiated scheduling mechanism in a 5G base station, which is applicable for gNB-1 and gNB-3 in the considered scenario. The differentiated scheduling of the Xcast traffic could be enforced in case of radio resource congestion situation, where due to mobility conditions or traffic variations impacting the shortage of radio resources available for delivering the data. The key assumption here is that the differentiated scheduling could be done by

isolating the Xcast users into a dedicated network slice. E.g., due to mobility the users might move to a worse radio condition requiring higher amount of radio resources for serving the same traffic, which requires dedicated scheduling for such users isolated from other unicast users. In case of congestion, when all the radio resources are utilized in a fully-loaded network, the data rates of the individual users need to be reduced. The throttling of traffic based on traffic type would affect Xcast traffic scheduled using multicast unfairly since the traffic is being delivered to more than one user as compared to unicast traffic.

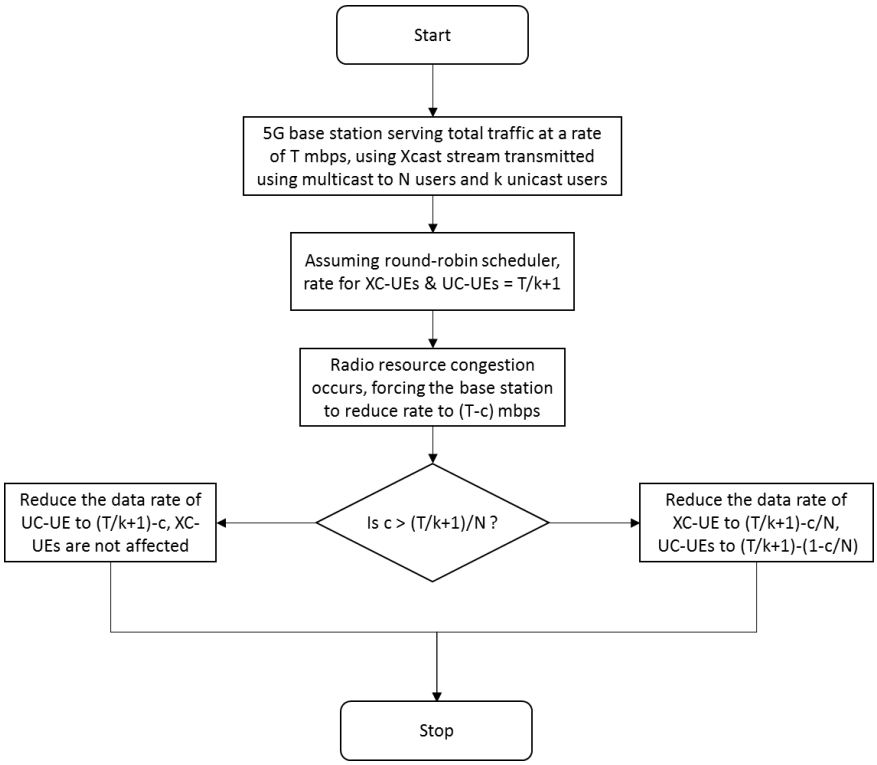


Figure 7-3 Flow diagram for the decision process of differentiated scheduling

Assume that the gNB is serving a total traffic of T mbps, and during radio congestion, the rate needs to be reduced to (T-c) mbps. In 5G-gNB1, there is one UC-UE receiving unicast data with a rate of T/2 mbps before congestion, and four XC-UEs receiving a single multicast stream at a rate of T/2 mbps. The differentiated 5G scheduler should consider the effective rate of XC-UEs to be (T/2)/4 during congestion for reducing the rates of the scheduled traffic. Thus, in the scenario considered here, if c < (T/2)/4, the Xcast stream would be left unthrottled, while the rate of UC-UE is reduced to (T/2)-c. This enables the 5G-gNB1 to serve the users at a reduced rate while not impacting the Xcast stream. If c > (T/2)/4, then the scheduler should reduce the rate of Xcast stream by c/4, and unicast stream by 3c/4. This generic decision-making process is indicated in the flow diagram shown in Figure 7-3. For 5G-gNB2, as mentioned earlier the XC-UE and UC-UE would be treated the same way, since the Xcast stream is delivered using unicast OTA. For 5G-gNB3, similar mechanisms described in the embodiment can be applied, with the consideration that the Xcast stream is received by two users, rather than four. The resultant impact on the traffic for the Xcast and unicast users due to the differentiated scheduling on the considered Xcast scenarios is as shown in Figure 7-4.

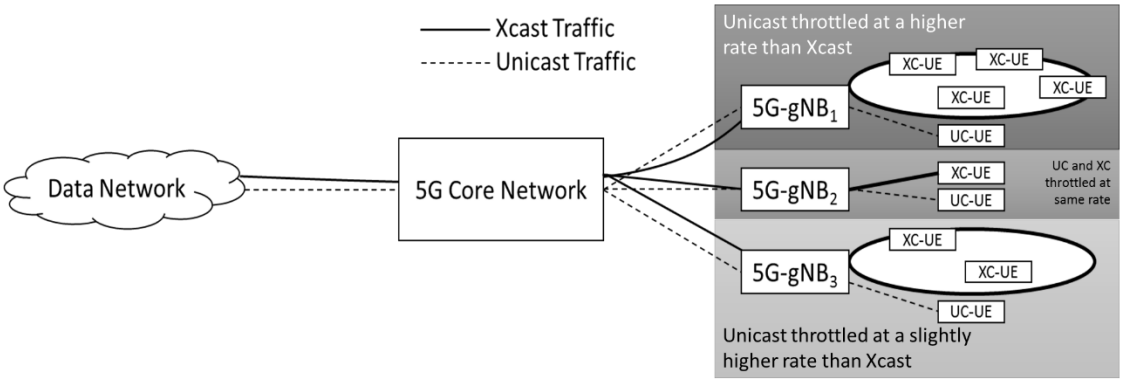


Figure 7-4 Resultant effect of differentiated scheduling on the different Xcast traffic scenarios.

7.3 Slicing RMA for 5G-Xcast Use Cases

5G-Xcast introduces RMA as the architectural and functional solution for broadcast and multicast (see subsection 3.1.1). It is based on C-RAN architecture and supports Single Cell PTM (SC-PTM) and Multiple Cell PTM (MC-PTM). Also based on the NGMN concept, in this subsection we describe how the RMA is sliced to support different 5G-Xcast use cases. That provides a framework to implement the network slicing in 5G-Xcast RAN, and sets the ground for future practical deployment as a primary option to provision and manage broadcast and multicast services.

5G-Xcast enhances gNB-CU-CP to make it capable of managing network slicing in RMA. The gNB-CU-CP (or CU-CP in short) maintains a list of NSST (Network Slice Subnet Template), a list of NSSI (Network Slice Subnet Instance), and optionally a list of NSI (Network Slice Instance). Figure 7-5 illustrates the management view of the 5G-Xcast RAN slicing. CU-CP is enhanced with slicing management and manages the lifecycle of all the network slices in the scope of RMA. Each network slice encompasses Virtualised Network Functions (VNFs) to meet the requirements of the specific slice. In the figure the VNFs are virtualised CUs and virtualised DUs, thus making each slice a fully functional NG-RAN. Also CU-CP needs to coordinate with the CN to ensure the end to end slice working properly, which is not shown in the figure.

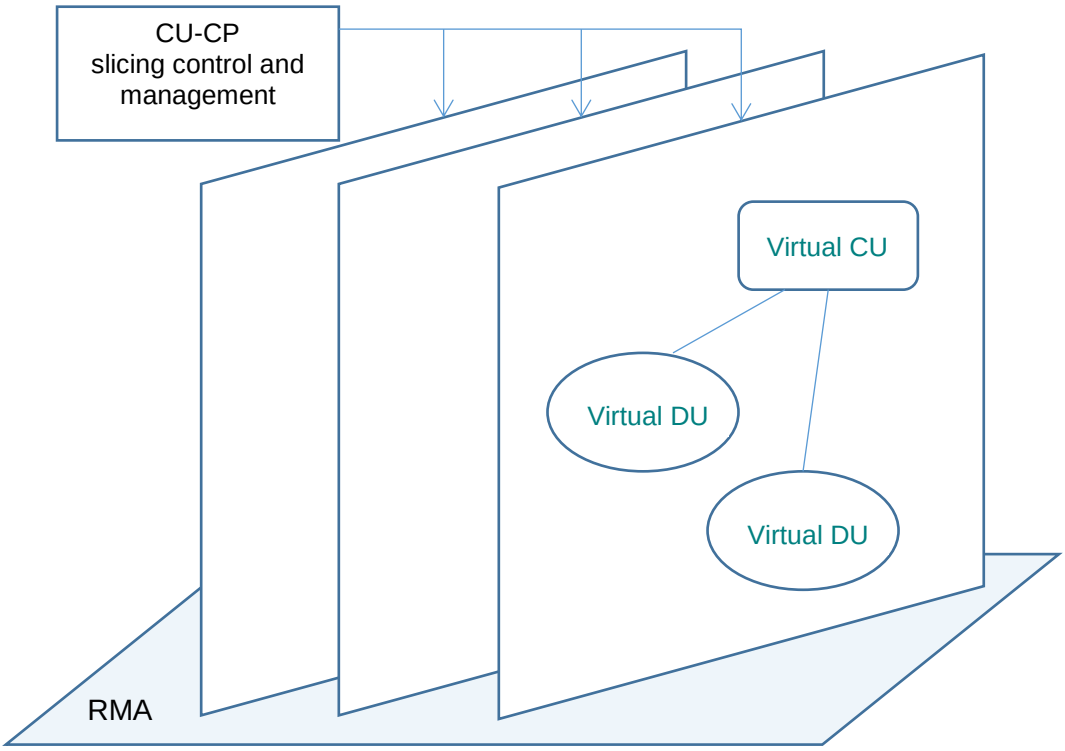


Figure 7-5: Management view of 5G-Xcast RAN slicing

In 5G-Xcast we adopt the layered concept of network slicing outlined by NGMN, where there are 3 layers: 1) Service Instance Layer, 2) Network Slice Instance Layer, and 3) Resource layer [60]. The VNFs need to be mapped to the physical Network Functions (NFs) which they rely on to execute the substantial networking functionality, as shown in Figure 7-6. The VNFs live on the slice instance layer while the PNFs are gathered on the resource layer. Note that VNFs of different slice instances can be mapped to the same PNF, but not different VNFs of the same slice instance. Alternatively, two VNF instances of the same type from the same slice could be mapped to the same PNF, e.g. mapping multiple DU-VNFs of the same slice to the same DU-PNF.

The isolation of the different network slice instances is enabled by partitioning of the PNFs, which will be elaborated further in section 7.4.

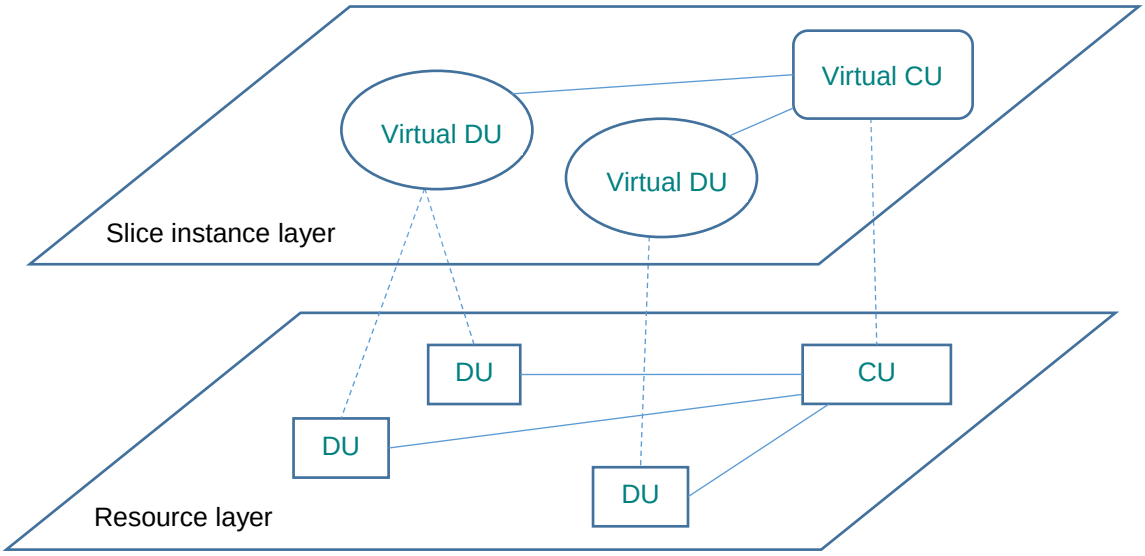


Figure 7-6: Resource mapping between slice instance layer and resource layer

Figure 7-7 shows an example how two network slice instances slice the network resources in RMA. The physical CU and DU in the RMA are shared by the virtual NFs in the slice instances 1 and 2 (which are Network Slice Subnet Instances discussed in section 7.4). In this way, the RMA is sliced into two parts, accommodating the two slice instances within the same RMA configuration, even if the RMA configurations can be also independent between slices.

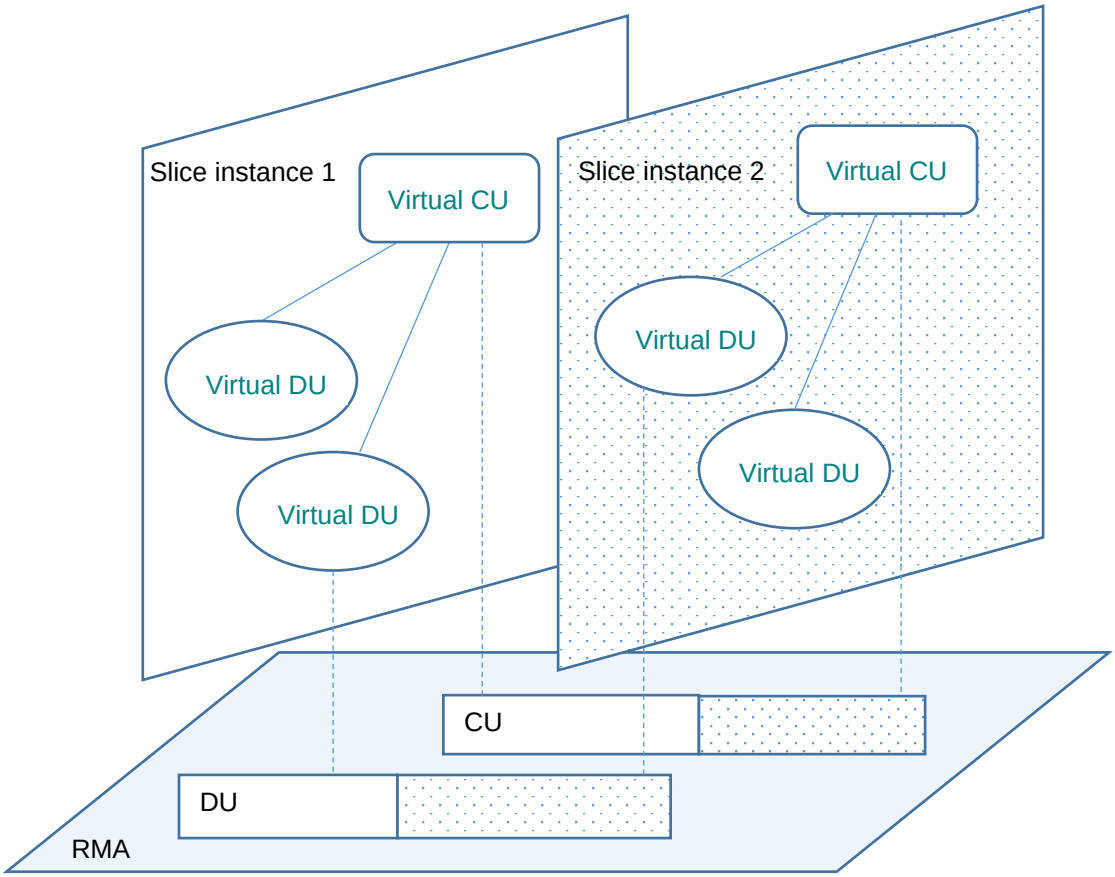


Figure 7-7: An example of two slice instances slicing shared resources in RMA

By harnessing network function virtualisation (NFV) and network softwarization, RMA can be sliced to facilitate the desired 5G network management solution. As pointed out in [48], it is not appropriate to define a pure multicast slice, as multicast is frequently mixed and tightly integrated with unicast to transport broadcast and multicast communication services. Further, there is a requirement [45] to allow deployment of a multicast solution that can seamlessly adapt between unicast and multicast transmission to maximize the efficiency of user radio and network resources. However, there is a need to define network slices for a category of broadcast and multicast services on the demand of Communications Service Provider (CSP) and according to the Service Level Agreement (SLA) signed with the Network Operator (NOP), or specifically Mobile Network Operator (MNO) for 5G networks. To facilitate the description, we suggest having a specific range Slice/Service Type (SST) reserved for 5G-Xcast service in 3GPP. As defined by 3GPP, the SST occupies 8 digits in an S-NSSAI (Single Network Slice Selection Assistance Information) comprising SST and SD (Slice Differentiator). To facilitate future investigation, experiment, and trials, we propose to reserve the following numbers as listed in Table *2* and Table *3*, which can be seen as an extension to the 3GPP standardised SST values in [7]. The corresponding QoS definitions, requirements and KPIs are given in [2][4].

5G-Xcast network slice category		SST range	
		Decimal	Binary
M&E	Media and Entertainment network slices	128 ~ 131	1000 00xx
PW	Public Warning network slices	132 ~ 133	1000 010x
Auto	Automotive network slices	134 ~ 135	1000 011x
IoT	Internet of Things network slices	136 ~ 137	1000 100x

Table 2: SST ranges for 5G-Xcast network slice categories

5G-Xcast network slice		SST	
		Decimal	Binary
M&E1	Hybrid broadcast service network slice	128	1000 0000
M&E2	Virtual / augmented reality broadcast network slice	129	1000 0001
M&E3	Remote live production network slice	130	1000 0010
PW1	Multimedia public warning alert network slice	132	1000 0100
Auto1	V2X broadcast service network slice	134	1000 0110
IoT1	Massive software and firmware updates network slice	136	1000 1000

Table 3: SST for 5G-Xcast network slices carrying 5G-Xcast communication services

5G-Xcast defines 6 use cases in another deliverable [2]. Based on the requirements and KPIs for those use cases we derive the slice characteristics for the correspondent SSTs, as specified in Table *4*.

Slice/Service Type	**Downlink data rate**	**End to end latency**	**Overall user density**	**UE speed**
M&E1	50 Mbps	< 50 ms	From 1000/km^2 to ≥10000/km^2	From stationary up to 500 km/h
M&E2	5 Gbps	< 7 ms	Several 100s /cell	From stationary up to 500 km/h
M&E3	9 Gbps	< a few ms	Up to 100 /site	Stationary
PW1	50 Mbps	< 50 ms, within 10 s	From 1000/km^2 to ≥10000/km^2	From stationary up to 500 km/h
Auto1	50 Mbps	< 5 ms	From 100/km^2 to 3000/km^2	From stationary up to 500 km/h
IoT1	50 Mbps	< 50 ms	≥10000/km^2	From stationary up to 500 km/h

Table 4: Service requirements for 5G-Xcast network slices

As each network slice is a logically separated network, the content synchronisation (see section 3.2) is not required between the slices.

Figure 7-8 illustrates the network slicing management for 5G-Xcast RAN with simplified signalling interaction between the 5G-Xcast RAN entities of a gNB. This is a framework for RAN to manage slice subnet, details like slice identifiers are reasonably omitted, as we don't elaborate signalling exchanges outside the RAN, which would involve UE and CN, where more details would get in including identifiers such as S-NSSAI. The RAN slicing management consists of two parts: (A) Network Slice Subnet (NSS) preparation; and (B) the lifecycle management (LCM) for Network Slice Subnetwork Instance (NSSI). Part (B) includes the basic NSSI commissioning, operation, and decommissioning. The application protocol E1AP is used to convey the signalling messages between CU-CP and CU-MC, and F1AP is used to convey signalling messages between CU-CP and DU. The UP traffic of broadcast and multicast is also depicted in the operation phase, where the uni-direction user traffic is transported with GTP-U. The procedures are described as follow:

(A) NSS preparation

A.1. CU-CP sends NSS Prepare request to CU-MC. The request includes the NSST description and all other specifications for the telecommunication resources to support the required broadcast and multicast slice.

A.2. CU-CP sends NSS Prepare request to DU. The request includes the NSST description and all other specifications for the telecommunication resources to support the required broadcast and multicast slice.

A.3. Upon the reception of A.1, CU-MC prepares network environment against the specification of the NSST, and sends Complete confirmation to CU-CP.

A.4. Upon the reception of A.2, DU prepares network environment against the specification of the NSST, and sends Complete confirmation to CU-CP.

(B) NSSI lifecycle management (LCM)

B.1. CU-CP sends NSSI Create request to CU-MC. The request includes the identifier of the NSSI and other specific information for creating the NSSI.

B.2. CU-CP sends NSSI Create request to DU. The request includes the identifier of the NSSI and other specific information for creating the NSSI.

B.3 Upon the reception of B.1, CU-MC reserves required telecommunication resources to the NSSI and sends Complete confirmation to CU-CP.

B.4 Upon the reception of B.2, DU reserves required telecommunication resources to the NSSI and sends Complete confirmation to CU-CP.

B.5. Upon the reception of B.1 and B.2, CU-CP sends NSSI Ready indication to CU-MC. Now the NSSI is commissioned.

B.6. Upon the reception of B.5, CU-MC can forward any broadcast and multicast user traffic to the destined DU(s). Now the NSSI enters operation.

B.7. During the operation, CU-CP may send Supervise request regularly or on demand to CU-MC to ensure the NSSI is operating properly.

B.8. Upon the reception of B.7, CU-MC should report the status of its NSSI NFs to CU-CP.

B.9. During the operation, CU-CP may send Supervise request regularly or on demand to DU to ensure the NSSI is operating properly.

B.10. Upon the reception of B.9, DU should report the status of its NSSI NFs to CU-CP.

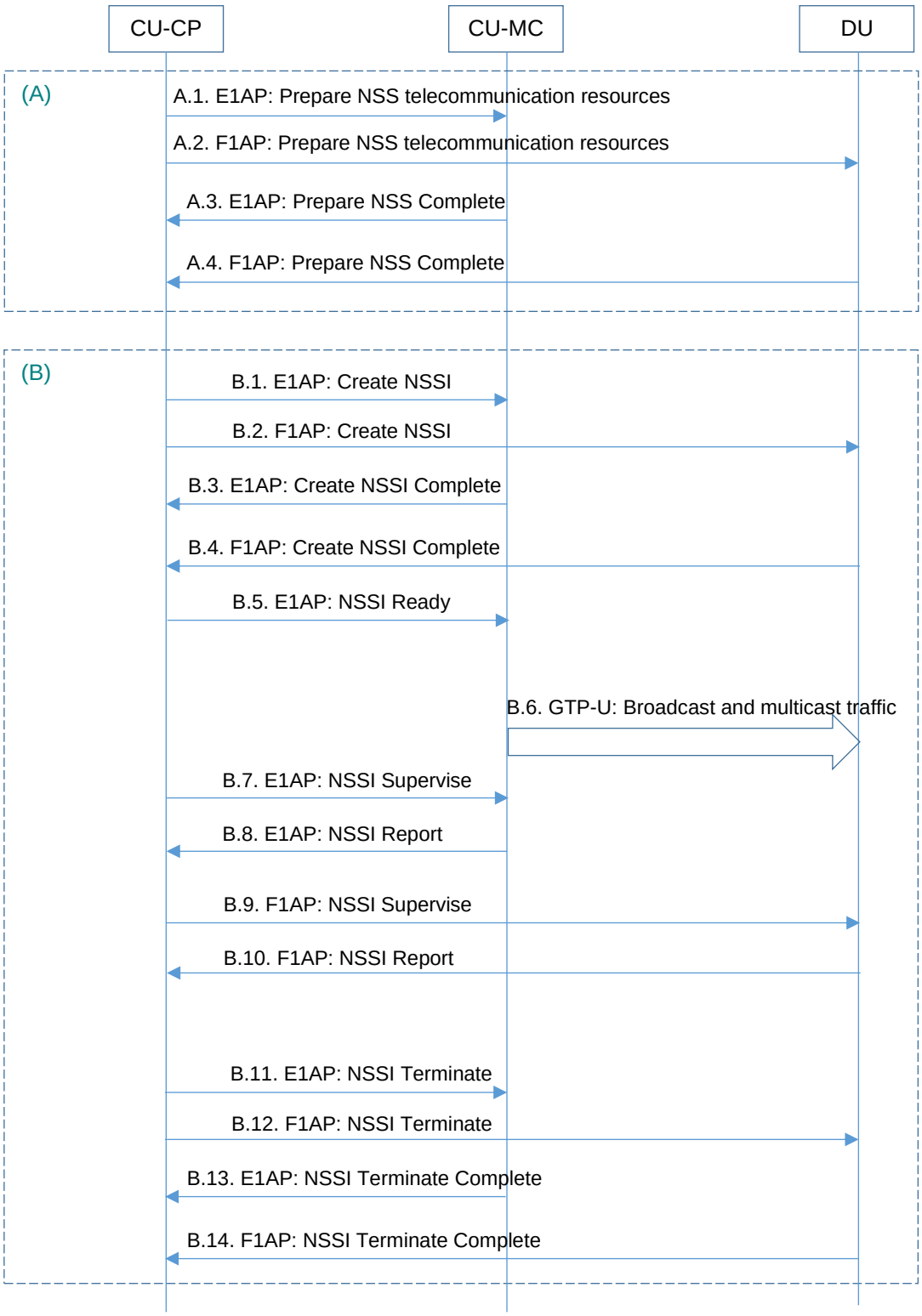
B.11. CU-CP may decommission the NSSI once there is such need, by sending a NSSI Terminate request to CU-MC, which includes the identifier of the NSSI.

B.12. At the same time of B.11, CU-CP sends a NSSI Terminate request to DU, which includes the identifier of the NSSI.

B.13. Upon the reception of B.11, CU-MC releases the telecommunication resources for the requested NSSI, and sends Complete confirmation to CU-CP.

B.14. Upon the reception of B.12, DU releases the telecommunication resources for the requested NSSI, and sends Complete confirmation to CU-CP.

For all RMA network slices or RMA shared by multiple slices, both Terrestrial Broadcast and Mixed Mode are supported, which is inherited from the 5G-Xcast RAN.



(A) NSS preparation (B) NSSI lifecycle management

Figure 7-8: Signalling interaction of network slicing management for 5G-Xcast RAN

7.4 Network Slice Subnet in 5G-Xcast RMA

A Network Slice Subnet (NSS) is “a representation of the management aspects of a set of Managed Functions and the required resources (e.g. compute, storage and networking resources)” [51]. The RAN part of the Network Slice is a Network Slice Subnet. Together with the Network Slice Subnet in the CN, they form a Network Slice [7]. In 5G-Xcast the concept of RAN sharing is at the centre of NSS in RMA.

As defined by 3GPP, a Network Slice Subnet Instance (NSSI) is “an instance of Network Slice Subnet representing the management aspects of a set of Managed Function instances and the used resources (e.g. compute, storage and networking resources)” [51]. Although 3GPP defines both shared and non-shared NSSIs, NSSIs in 5G-Xcast are non-shared in order to provide management convenience, more network isolation, and ensure the resource availability for the CSC (Communication Service Customer), especially when provided as NSaaS (Network Slice as a Service). This also complies with the business rationale of network slicing, where a CSP normally will not further share their own network with other CSPs unless some Virtual Network Functions (VNFs) and associated resources are reallocated and/or another NSSI is created for that purpose. The CSP could also use its infrastructure to serve multiple verticals by using dedicated slices run on the same infrastructure. The network sharing happens underneath at the physical network, through real time network slicing scaling where needed. In other words, network sharing happens at the Resource Layer managed by NOP (Network Operator. Here we follow the concept specified by 3GPP [51], view Network Slices as NOP internals, and do not use the term CSP), rather than at the Slice Instance Layer managed by CSP (for the definition of the network slicing layered structure see [60]), although NOP and CSP can be taken as a same role in a practical deployment.

According to the requirement set by 3GPP, “the 5G system shall support scaling of a network slice, i.e., adaptation of its capacity” [55]. “The new RAT shall support dynamic adjustment of the Multicast/Broadcast area based on e.g. the user distribution or service requirements” [45]. 5G-Xcast introduced dynamic network slice management with a flexible scaling to address those requirements.

7.4.1 Telecommunication resource pool

5G-Xcast puts the telecommunication resources in the RAN to a virtual pool, including mainly managed NFs (VNFs and PNFs), management functions and network resources. That is particularly compliant with the C-RAN architecture where all gNB-CU functions are already in a centralised pool. The resource pool stands for the entire capacity of the 5G-Xcast RAN and the available capacity for all the NSS. Every time a NSSI is created, the resources needed by the NSSI are reserved for the NSSI and the RAN is sliced. Figure 7-9 illustrates an example of telecommunication resource pool in 5G-Xcast RAN. The resources are mainly managed NFs. There are n NSSIs already created. For each NSSI there are a number of NFs reserved and ready to be employed once there is active usage of that NSSI. The available NFs could be gathered into an NSSI 0 which is not shown for the sake of managerial convenience. Note that the NFs, which are virtual on the slice instance layer, are not shared between NSSIs once reserved.

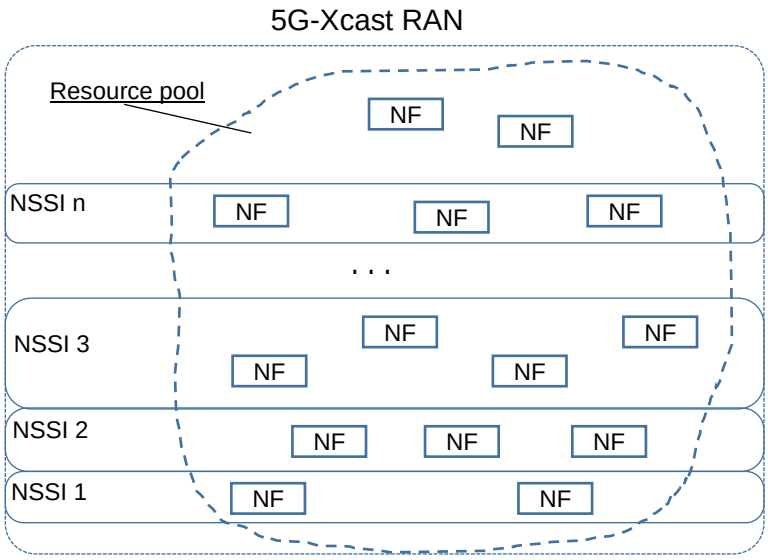


Figure 7-9: Telecommunication resource pool in 5G-Xcast RAN

7.4.2 Inter-slice isolation

Inter-slice isolation is an important concept for 5G-Xcast NSS. It guarantees resources for NSSIs and consequently the QoS that can be provided to the NSSI users.

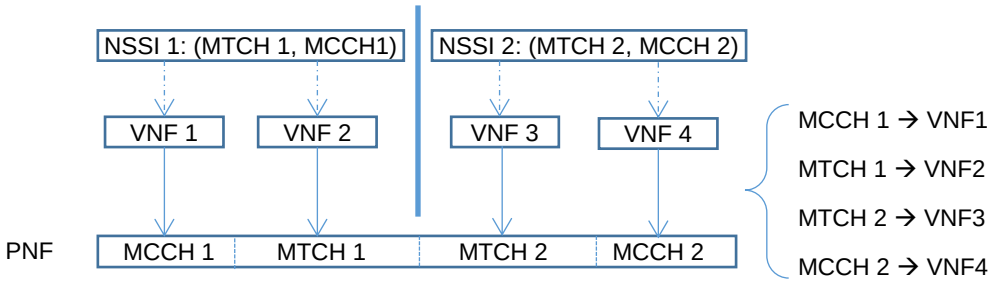


Figure 7-10: Isolated VNFs mapping to PNF

Figure 7-10 illustrates an example of inter-slice isolation. NSSI 1 encompasses VNF 1 and VNF 2; NSSI 2 encompasses VNF 3 and VNF 4. VNF 1 is a virtualisation of logical channel MCCH 1; VNF 2 is a virtualisation of MTCH 1; VNF 3 is a virtualisation of MTCH 2; VNF 4 is a virtualisation of MCCH 2. The logical channels are PNFs, shared among the VNFs. As a matter of fact, NSS 1 harnesses MTCH 1 and MCCH 1, and NSS 2 harnesses MTCH 2 and MCCH 2.

7.4.3 Flexible slice scaling

In 5G-Xcast we introduce a flexible slice adjustment method which allows dynamic scaling during NSSI operation. Whenever needed, CU-CP may upscale or downscale an NSSI by reserve more or fewer VNFs for it. In Figure 7-10, if the multicast traffic in NSSI 1 reduces to nil for some specific period, while traffic in NSSI 2 is foreseen to increase, there is a need to downscale NSSI 1 and upscale NSSI 2. By moving the isolation line towards left, NSSI 1 is downscaled through moving VNF 2 to NSSI 2, thus

increasing the capacity of the latter. Figure 7-11 shows an example how NSSI 1 is downscaled while NSSI 2 is upscaled by “sliding” the boundary between them. Note that all those operation are in a software manner by re-assign the reserve from one NSSI to another, without a real change to the physical layer resources.

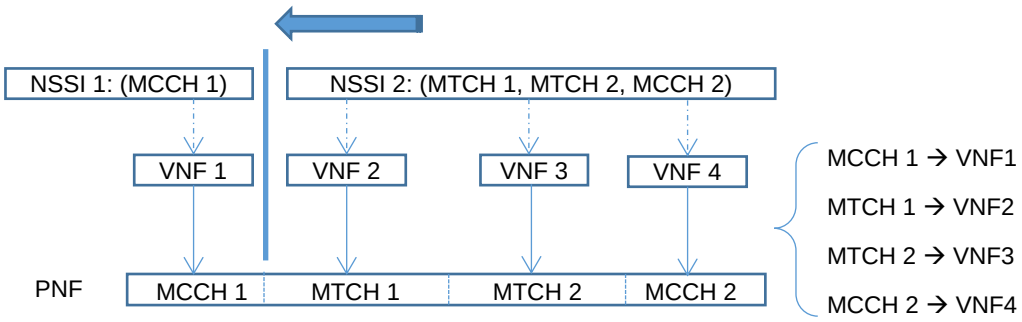


Figure 7-11: NSS scaling in 5G-Xcast

7.5 Summary

In this section we focus on the important aspect how 5G-Xcast RAN supports network slicing. First we give a general scenario where unicast, multicast and broadcast traffic are transmitted in a mixed mode. In the case of network congestion at the base station, we selectively throttle different types of the traffic during the traffic scheduling. The key assumption here is that the differentiated scheduling could be done by isolating the broadcast and multicast users into a dedicated network slice. Following that, we propose a slicing methodology based on 5G-Xcast use cases specified in the preceding work on 5G-Xcast user cases, requirements and KPIs. The proposal is based on NGMN concepts and a further enhancement to standard 3GPP CU-DU RAN architecture. The roles played by Network Function Virtualisation and network softwarization are also explained. To facilitate further research and development, we suggest reserving a few designated Slice/Service Type numbers, which can be seen as an extension to the SST numbers currently defined by 3GPP. At last, we give more details on Network Slice Subnet design in the 5G-Xcast RAN. The innovative concepts of telecommunication resource pool, inter-slice isolation, and flexible slice scaling are proposed as the core of 5G-Xcast RAN slicing.

8 Summary and Conclusions

8.1 Introduction

The first release of NR focused on unicast communications in the core network and PTP transmissions in the RAN. As the NR evolves to new services and vertical sectors, such as automotive, IoT, M&E and PWS, the network needs to deliver a common content simultaneously to many users. The flexible RAN architecture will be able to select the most effective transmission method between unicast, multicast and broadcast, also considering the number of users interested in receiving the content, user geographical distribution and type of service. In 5G-Xcast RAN architecture these requirements were considered and taken as the design principle.

3GPP Rel-15 specifications have extended the stand-alone base station architecture towards flexible cloud-RAN based functionality where the split between CUs and DUs enables dynamic adaptation of QoS functions. These architecture enhancements have been utilized in the innovative RAN architecture designed for 5G-Xcast.

5G-Xcast RAN architecture will be aware of the user's interest to receive IP multicast data, as well as the same architecture allows deployment of broadcast services towards ROM devices and services such as PWS. The developed concepts of dynamic RMA and flexible TB-SA allow delivery of multicast/broadcast services wherever needed without the fixed deployment on top of the existing RAN.

The multicast/broadcast services over multiple cells require content synchronization to share the same content among multiple users. 5G-Xcast RAN architecture supports the multi-cell and SFN transmission using NG-RAN based synchronization method fulfilling the QoS targets defined for the traffic flows.

In this deliverable, the 5G unicast architecture was evolved to enable efficient selection between unicast and multicast/broadcast transmission modes resulting a RAN architecture design which can adapt to a wide variety of requirements and QoS targets defined according to user demand and traffic flows and also network operators and service providers.

8.2 Evaluation of 5G-Xcast RAN Architecture against 5G Broadcast/Multicast requirements

3GPP specification 38.913 [45] sets the requirements for 5G Broadcast/Multicast systems. Following Table *5* compares the developed RAN Architecture solutions against these requirements for 5G Broadcast/Multicast systems

Broadcast/Multicast Requirement (*RAN architecture related*)	**5G-Xcast RAN Architecture solution (*section reference*)**
The new RAT shall support *existing Multicast/Broadcast services* (e.g. download, streaming, group communication, TV, etc.) and *new services* (e.g. V2X, etc).	• The overall 5G-Xcast RAN architecture supporting multicast/broadcast and vertical segments. • PDU Session Modification for multicast resource allocation and multicast bearer selection. *Sec. 3.1 Functional entities to support Point-to-*

	Multipoint in 5G architecture, *Sec. 6.1 5G-Xcast deployment scenarios*
The new RAT shall support *dynamic adjustment of the Multicast/Broadcast area* based on e.g. the user distribution or service requirements.	• RAN based Multicast Area (RMA) to allow dynamic multicast area along with user distribution, requested service, UE unicast activity, mobility and connectivity. • RMA procedures to create and modify RMA. *Sec. 5.2 RAN Multicast Area procedures*
The new RAT shall support *static and dynamic resource allocation between Multicast/Broadcast and unicast*; the new RAT shall in particular allow support of up to *100% of DL resources for Multicast/Broadcast* (100% meaning a dedicated MBMS carrier).	• RAN based content synchronization supporting Transparent Multicast and Terrestrial Broadcast. • Support for SFN with TB-SA and RAN based synchronization using gNB-CU-MC enabling synchronized transmission within the gNB-DUs. • Multicast deployment using Cloud RAN for seamless unicast/multicast bearer selection. *Sec. 3.2 RAN Synchronization for network entities,* *Sec. 5.2 RAN Multicast Area procedures,* *Sec. 6.1.3 Cloud-RAN based deployment,* *Sec. 6.1.5 Terrestrial Broadcast deployment*
The new RAT shall support Multicast/Broadcast *network sharing between multiple participating MNOs*, including the case of a dedicated MBMS network.	• Support for SFN with TB-SA and RAN based synchronization. • RAN slicing using RAN Multicast Area *Sec. 6.1.5 Terrestrial Broadcast deployment* *Sec. 7 5G-Xcast RAN Slicing*
The new RAT shall make it possible to *cover large geographical areas up to the size of an entire country in SFN mode with network synchronization* and shall allow cell radii of up to 100 km if required to facilitate that objective. It shall also support local, regional and national broadcast areas.	• Synchronized multicast content transmission using gNB-CU-MC which enforces the synchronized over the air transmission of multicast/broadcast traffic within the gNB-DUs. • Support for SFN with RAN based synchronization. *Sec. 3.2.3 Synchronized multicast content transmission over Single Frequency Networks,* *Sec. 5.2.5 RAN Multicast Area Procedures for Terrestrial Broadcast,* *Sec. 6.1.5 Terrestrial Broadcast deployment*
The new RAT shall support Multicast/Broadcast *services for fixed, portable and mobile UEs*. Mobility up to 250 km/h shall be supported.	• The overall 5G-Xcast RAN architecture. • Support for ROM and PDU Session Modification for multicast resource allocation

	and multicast bearer selection. *Sec. 3.1 Functional entities to support Point-to-Multipoint in 5G architecture,* *Sec. 3.2 RAN Synchronization for network entities,* *Sec. 6.1.5 Terrestrial Broadcast deployment*
The new RAT shall *leverage usage of RAN equipment* (hard- and software) including e.g. multi-antenna capabilities (e.g. MIMO) to improve Multicast/Broadcast capacity and reliability.	• The overall 5G-Xcast RAN architecture, e.g. based on the design principle of maximizing the architectural commonality with unicast. *Sec. 3.1 Functional entities to support Point-to-Multipoint in 5G architecture*
The new RAT shall *support Multicast/Broadcast services for mMTC devices*.	• The overall 5G-Xcast RAN architecture is designed to support all kind of devices, e.g. devices targeting for IP multicast services with TX/RX capability and ROM mode devices. *Sec. 3.1 Functional entities to support Point-to-Multipoint in 5G architecture,* *Sec. 6.1.3 Cloud-RAN based deployment,* *Sec. 6.1.5 Terrestrial Broadcast deployment*

Table 5: 5G-Xcast RAN Architecture solutions compared to 3GPP specification 38.913 requirements for 5G Broadcast/Multicast systems.

8.3 Summary

This deliverable describes the overall 5G-Xcast RAN logical architecture solution and the related details. The RAN architecture was shown to fulfil the 5G-Xcast use case specific requirements depicted in deliverable D2.1 [2], address the technical developments described in D2.2 [3] and cover the generic architectural requirements listed in 3GPP TS 38.913 [45].

The NR Rel-15 RAN unicast architecture was assumed as a basis for the 5G-Xcast architecture design. The developed 5G-Xcast RAN architecture extends the NR unicast architecture allowing dynamic deployment of 5G-Xcast RAN where the new RAT supports dynamic adjustment of the Multicast/Broadcast geographical area based on e.g. the user distribution or service requirements. The new 5G-Xcast RAN architecture can cover large geographical areas up to the size of an entire country in SFN mode with content synchronization for SFN transmission. Developed RAN Multicast Area and RAN based synchronization solutions can support local, regional and national multicast/broadcast areas. The support for dynamic geographical areas is enabled with support of concurrent delivery of both unicast and Multicast/Broadcast services to the users, as well as support for efficient multiplexing with unicast transmissions via seamless data bearer selection.

A Appendices

A.1 5G-Xcast system architecture alternatives [1]

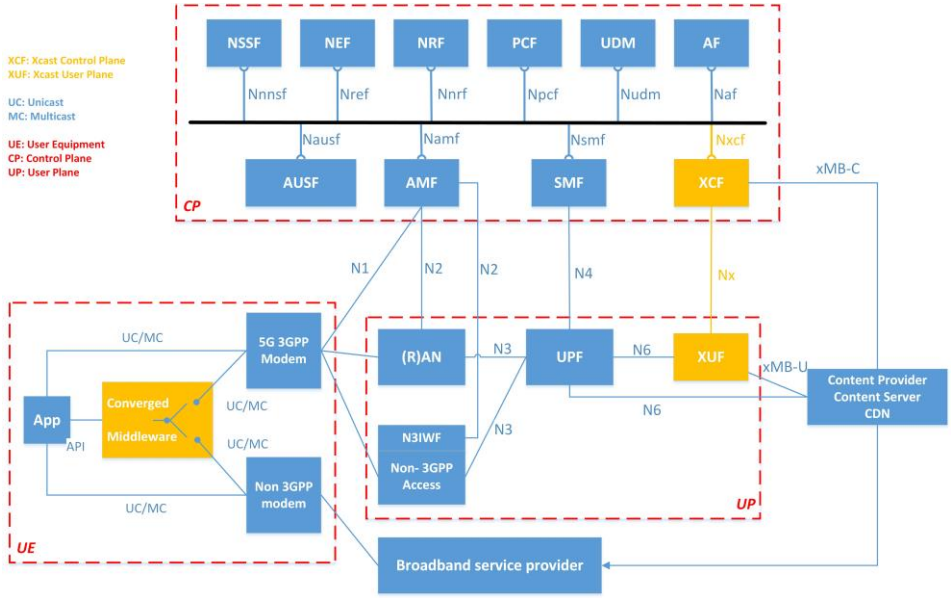


Figure A-1: 5G-Xcast system architecture alternative 1

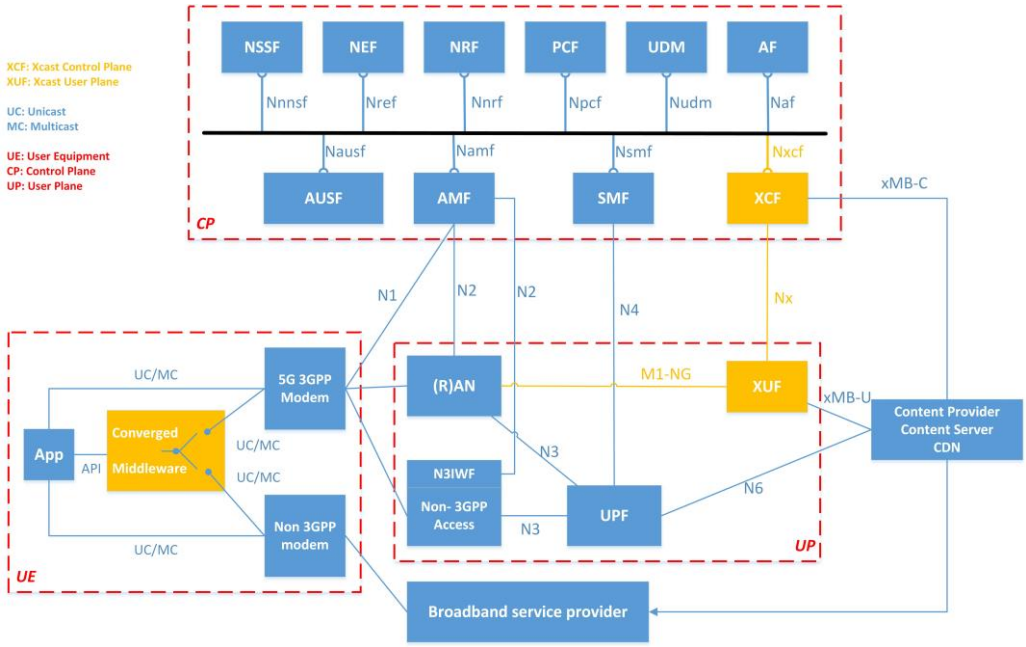


Figure A-2: 5G-Xcast system architecture alternative 2

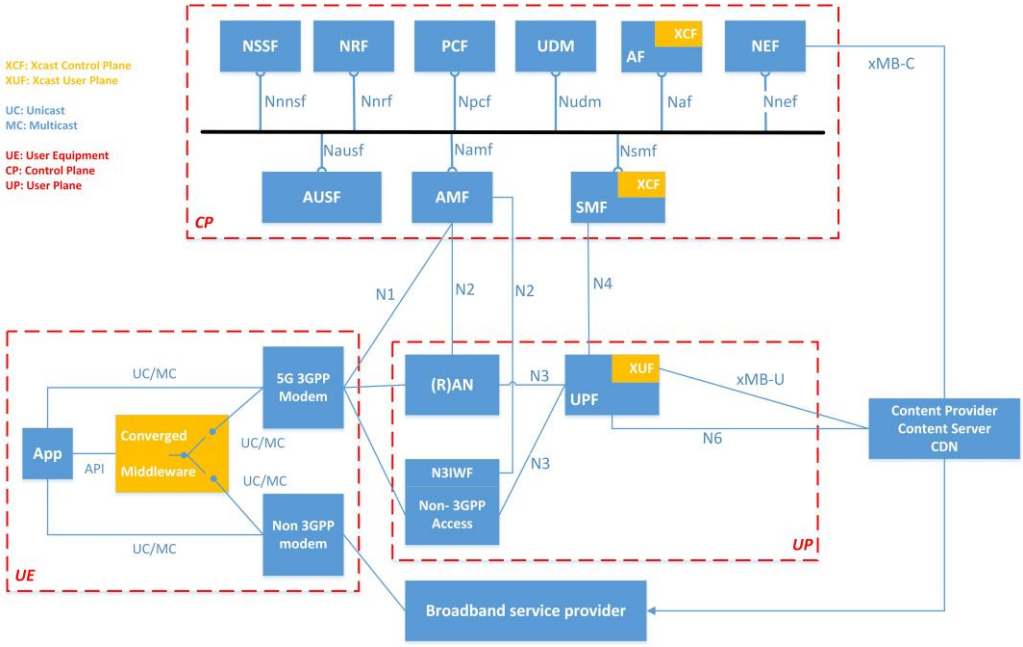


Figure A-3: 5G-Xcast system architecture alternative 3

A.2 Protocol stack across 5G system (SDAP, PDCP, RLC, MAC, PHY)

5G-Xcast RAN will use the NR protocol stack as specified in [11]. That will ensure the smooth evolution of and extension to the current 5G architecture. The UP and CP stacks are represented respectively in Figure A-4 and Figure A-5.

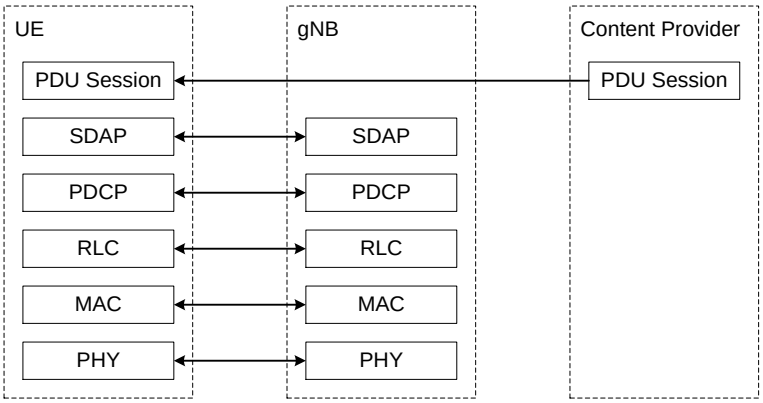


Figure A-4: 5G-Xcast UP protocol stack

In the UP stack, there are sublayers SDAP, PDCP, RLC and MAC in L2 to transfer the upper layer PDU Session. The broadcast and multicast PDU Session originates in the Data Network (DN), where specifically for 5G-Xcast the network entity functionally is a broadcast and multicast Content Provider.

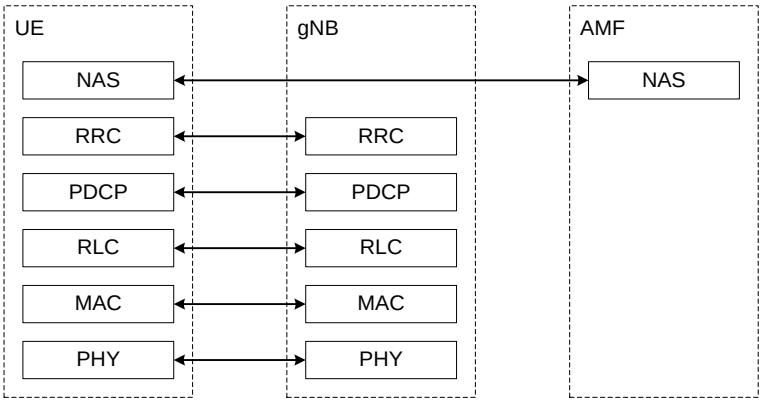


Figure A-5: 5G-Xcast CP protocol stack

In the CP stack, there are sublayers PDCP, RLC and MAC in L2 to transfer RRC signalling.

A.2.1 SDAP

5G-Xcast SDAP is on the UP and is an enhancement to the 3GPP specifications [11] and [15] to support broadcast and multicast QoS flows. SDAP is responsible for mapping QoS flows to multicast DRBs in 5G-Xcast and marking QoS flow ID (QFI) in the DL broadcast and multicast packets. The SDAP supports mapping of one or more

QoS flows onto one multicast DRB. The interlayer protocol structure, which is an enhancement to [15], is shown in Figure A-6. The SDAP entities are located in the SDAP sublayer in the protocol stack and an SDAP entity is configured for each individual PDU session. When an SDAP entity receives an SDAP SDU from upper layers, it constructs the corresponding SDAP PDU and submits it to lower layers. On the other hand, when an SDAP entity receives an SDAP PDU from lower layers, it retrieves the corresponding SDAP SDU and delivers it to upper layers.

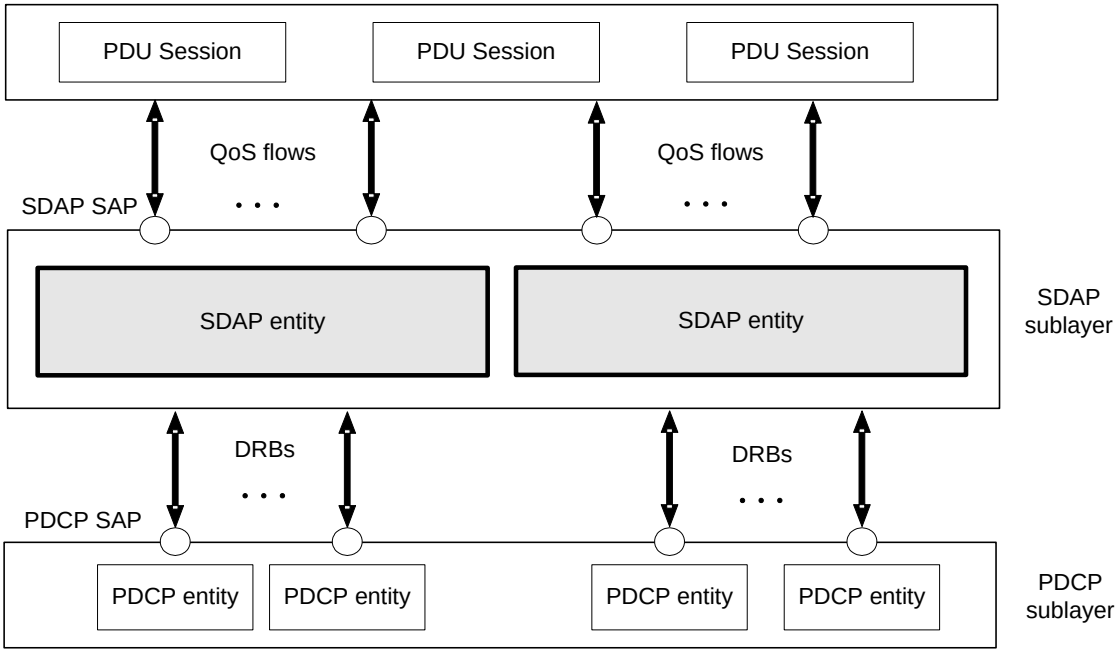


Figure A-6: 5G-Xcast SDAP protocol structure

A.2.2 PDCP

5G-Xcast PDCP provides DRBs and SRBs to its upper layers. It performs sequence numbering, reordering and duplicate detection, retransmission, header compression and decompression, ciphering, deciphering and integrity protection, and so on [11][14]. The PDCP will use lower layer RLC UM entities for broadcast and multicast service delivering without feedback for ARQ, link adaptation and HARQ. On the other hand, PDCP uses lower layer RLC AM entities for MBMS delivering with feedback signalling for data unit re-transmission requests. As specified in [14], the PDCP interlayer protocol structure is shown as in Figure A-7. Herein, a PDCP entity is configured to carry the data of one multicast DRB from the SDAP layer. On the other hand, except SRB0, one PDCP entity is configured to carry control data of one multicast SRB from RRC layer. 5G-Xcast PDCP functions are considered without precluding the option to operate in transparent mode as in the case of existing LTE-A eMBMS systems [8].

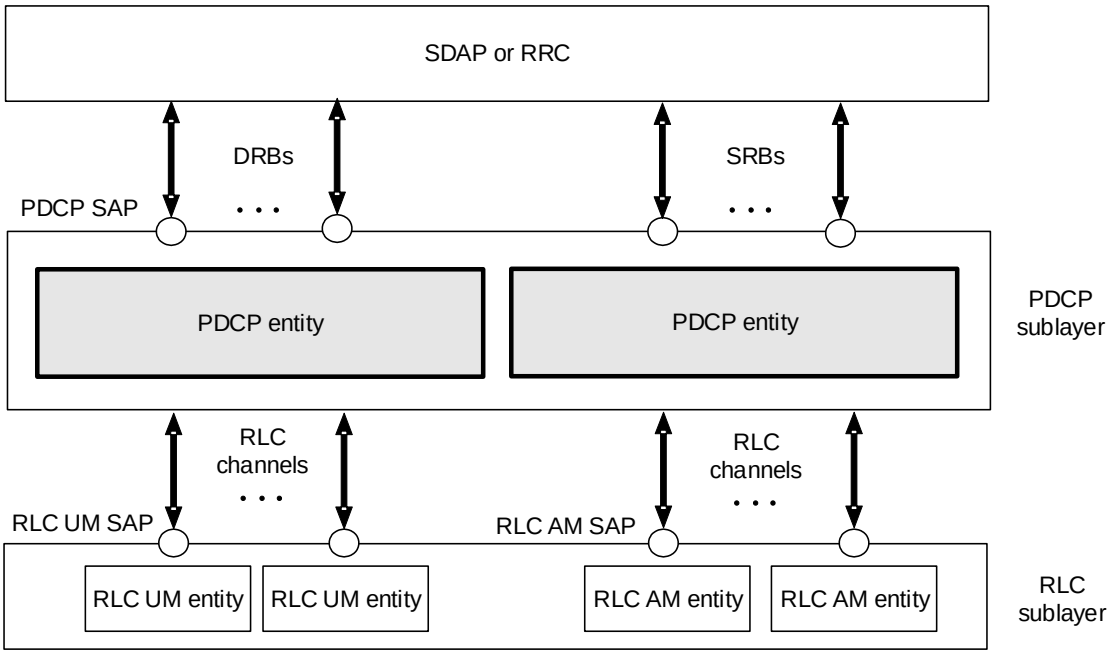


Figure A-7: 5G-Xcast PDCP protocol structure

A.2.3 RLC

In line with the 5G NR specification [11], 5G-Xcast RLC services and functions include sequence numbering independent of the one in PDCP, segmentation, re-segmentation and reassembly, discard of RLC SDUs. In accordance with the 5G NR specification in [13], RLC provides TM, UM and AM RLC entities to receive from its upper layers. A TM RLC entity is transparent to the PDUs that pass through it (no functions and no headers), and it is used for reception/delivery of control data via CCCH, BCCH and PCCH logical channels. A UM RLC entity carries either one complete broadcast and multicast RLC SDU or a RLC SDU segment, and it performs segmentation functions and adds/removes relevant RLC header information. UM RLC entity is used for reception/delivery of data or control broadcast and multicast RLC SDUs via SC-MTCH, MTCH, SC-MCCH and MCCH logical channels. On the other hand, an AM RLC entity carries either one complete broadcast and multicast RLC SDU and a RLC SDU segment, and it performs segmentation, re-segmentation and adds/removes the relevant RLC headers. Besides, AM RLC entity includes functions for protocol data unit re-transmission, duplicate detection, protocol error detection. The 5G-Xcast RLC interlayer protocol structure, specifically for transporting MBMS, is shown as in Figure A-8.

In 5G-Xcast, both RLC UM and RLC AM modes are considered for delivery of data or control broadcast and multicast RLC SDUs via SC-MTCH, MTCH SC-MCCH and MCCH logical channels. For terrestrial broadcast with receive only devices, RLC UM mode can be used. On the other hand, for mixed-mode broadcast/multicast, RLC AM can be an option used to provide highly re-liable and efficient delivery of broadcast and multicast service data. Herein, an efficient re-transmission scheme suitable for broadcast and multicast applications should be used.

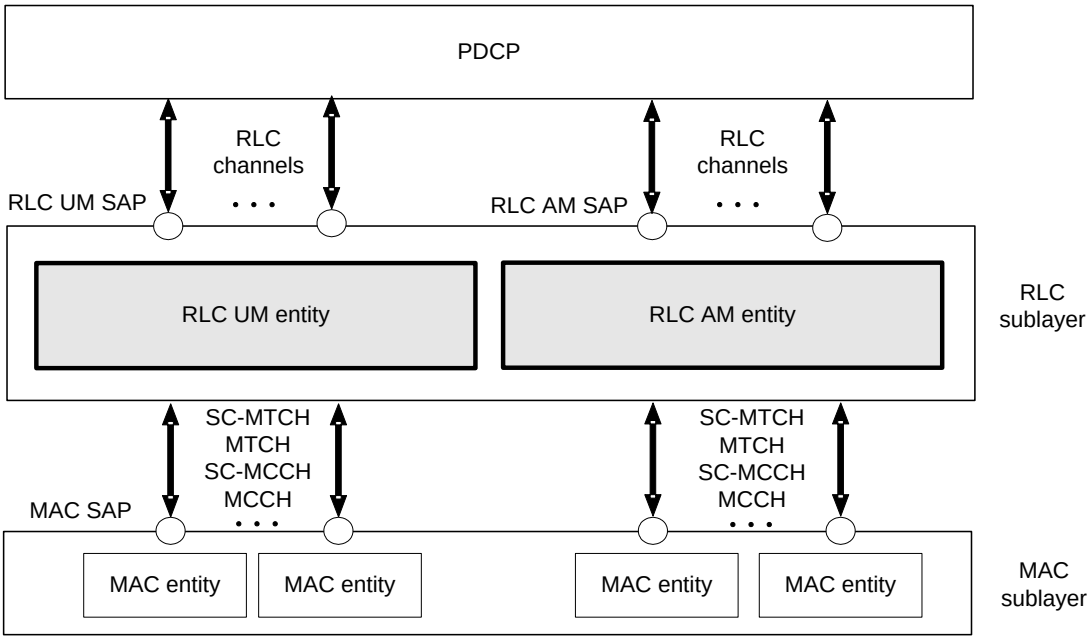


Figure A-8: 5G-Xcast RLC protocol structure

A.2.4 MAC

Developed based on 3GPP specifications for E-UTRA and NR [10][12], 5G-Xcast MAC provides various logical channels over which to transport upper layer UP traffic data and upper layer CP information, and the channels are accordingly classified into Traffic Channels and Control Channels. To support broadcast and multicast, the Control Channels used in 5G-Xcast are BCCH, SC-MCCH, MCCH and CCCH. The Traffic Channels are SC-MTCH and MTCH. MAC main services and functions include

- Broadcast and multicast service identification,
- Transport format selection,
- Mapping between logical channels and transport channels,
- Priority handling between different UEs by dynamic scheduling,
- Priority handling between different logical channels for one UE,
- Scheduling information reporting,
- Error correction through HARQ,
- Multiplexing and demultiplexing of MAC SDUs, and
- Padding.

The MAC interlayer protocol structure is shown as in Figure A-9.

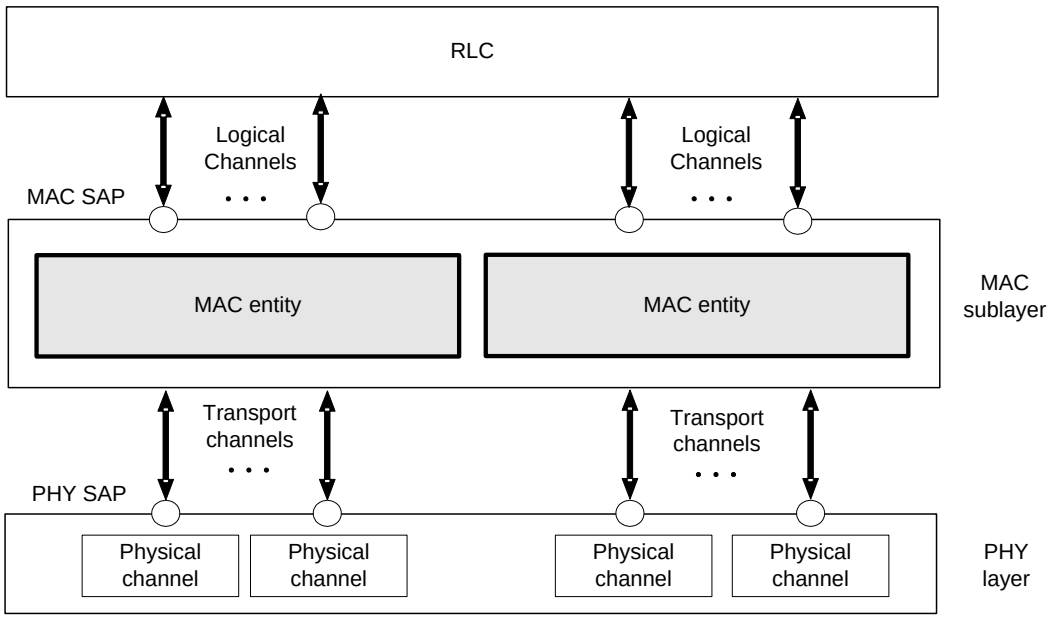


Figure A-9: 5G-Xcast MAC protocol structure

A.2.5 RRC

5G-Xcast RRC is on CP and responsible for transporting signalling between the UE and the 5G RAN, for both NAS and AS. RRC is responsible to manage its local SRBs and DRBs, and the RRC signalling to its communicating counterpart is transported over SRBs. The RRC in the protocol structure is shown in Figure A-10.

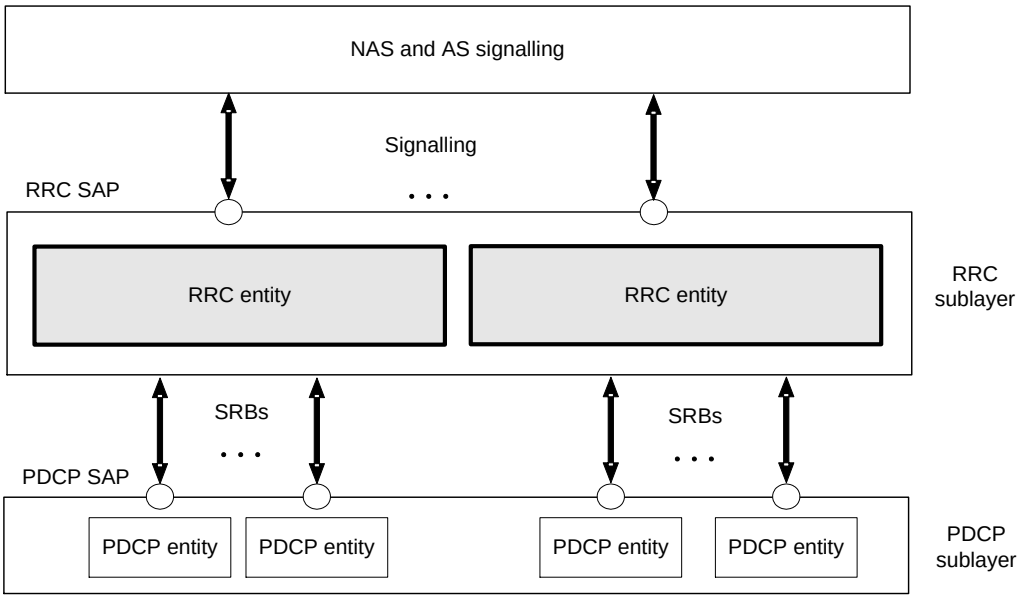


Figure A-10: 5G-Xcast RRC in the protocol structure

A.3 Applications of RAN Logical Architecture for Immersive Content Delivery

A.3.1 Introduction to Immersive Content Delivery

Here we will discuss the potential applications for the RAN logical architecture for 5G-Xcast presented in this deliverable, especially in the context of immersive content delivery. The mass delivery of immersive content such as virtual or augmented reality (VR / AR) is one of the use cases considered in [2], which raises unique challenges for multicast / broadcast as compared to unicast, as discussed in [75]. Some of the key challenges identified, especially in the context of high-quality media delivery, could be summarized as:

- The peak data rates currently possible for multicast / broadcast transmissions which limits the delivery of immersive content with possible requirements of gigabits per second
- Need for higher or near-unicast spectral efficiency
- Agile and flexible end-to-end architecture with minimal latency inducing elements (due to the latency requirements of such content)
- Lack of system bandwidth and available licensed spectrum for efficiently delivering highly immersive content

Various challenges described here requires inherent solutions within the 5G-Xcast system – such as enhanced air interface design for improving spectral efficiency, which has been evaluated in deliverable D3.2 [3], and simplified / agile end-to-end architecture in D4.1 [1]. Here, we will discuss some of the solutions that were developed as applications of the 5G-Xcast RAN architecture, to address the key challenges faced by the mass delivery of immersive content. The focus is mainly on optimizing the broadcast coverage area while minimizing feedback, optimizing the data transmitted over-the-air and enabling transmissions over unlicensed spectrum in order to improve the availability of spectrum for transmitting immersive content.

A.3.2 D2D-Augmented X-casting

For the scenario with indoor broadcast of immersive content, we consider the possibility to optimize the coverage area for the broadcast transmissions with the usage of device-to-device (D2D) communications for augmenting the coverage area of the primary broadcast transmitter. This enables the network to optimize the radio parameters used for the broadcast transmission from the base station for a reduced coverage area, thereby minimizing the radio resources utilized for the transmission. Here the key assumption is that the 5G UEs can augment the coverage area using D2D broadcast. The basic operation of the mechanism is as shown in the figure below [76], where the 5G-gNB is optimized to cover a subset of users (called VR broadcast / VR-BC) with users within the edge of the coverage area of such transmissions augment the coverage using D2D broadcast (D2D-BC) transmissions.

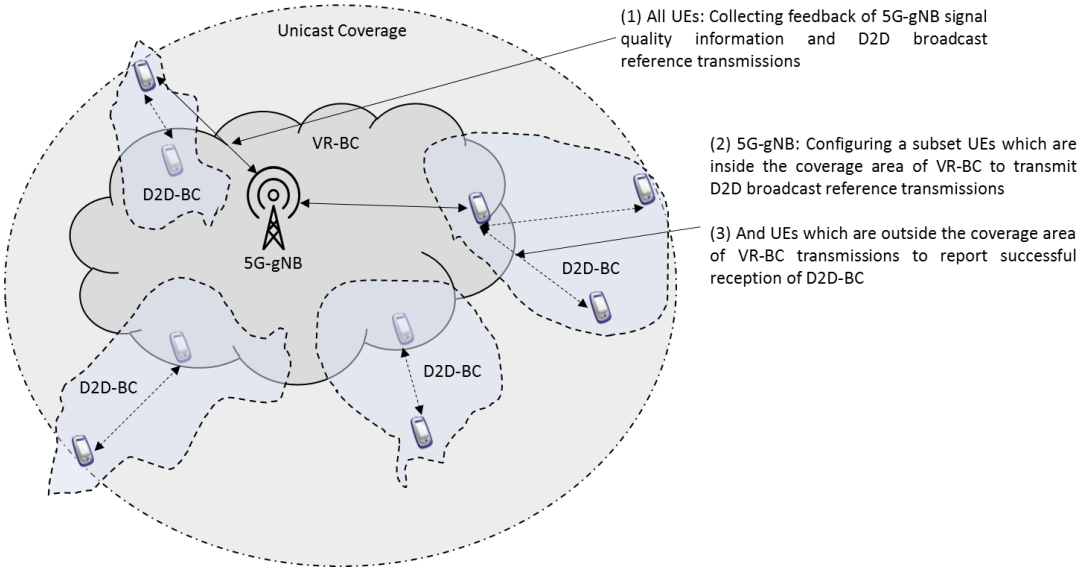


Figure A-11: Configurations for enabling D2D-augmented broadcast [76]

The configurations for enabling the D2D-augmented broadcast is as shown in Figure A-11. Here the 5G-gNB configures UEs to report the received signal quality information for the signals from the base station along with the D2D broadcast reference signal transmissions. Based on this information, the gNB could optimize the broadcast transmissions for an optimized coverage area which would be smaller than the one required to cover all the UEs. The gNB can configure the devices in the cell-edge of the optimized transmissions to resend the data using D2D-BC, thereby providing coverage for all the users with an optimized amount of resources. Based on the evaluations presented in [76], it is shown that the total system bandwidth required for broadcasting fully-immersive VR content is significantly reduced by over two times, assuming that the same spectral resources are reused for D2D-BC and VR-BC. The feedback within the system could be minimized using pre-configurations, in terms of the radio parameters used for the main VR-BC transmissions and the threshold for UEs to initiate D2D-BC transmissions.

A.3.3 Usage of Unlicensed Spectrum

The significant high data rate requirement of immersive content requires the availability of large amounts of spectrum, which needs to support multicast / broadcast transmissions in order to transport the content efficiently to the end users. The support for unlicensed spectrum would also enable new use cases and business models for 5G-Xcast in terms of new service offerings and deployment scenarios. The enhancements required for enabling such transmissions using a modified synchronization protocol is considered in [77], using LTE eMBMS system as the baseline. The proposed enhancement is as shown in the Figure A-12, where a system with multiple base stations coordinates the multicast/broadcast transmission over unlicensed spectrum while experiencing partial interference which prevents transmissions from one of the base stations during a limited period. Based on the evaluations presented in [77], it is also shown that by coordinating the channel sensing and data transmission time periods, the system can significantly improve the spectral

efficiency of the multicast / broadcast transmissions, as compared to all the base stations operating independently.

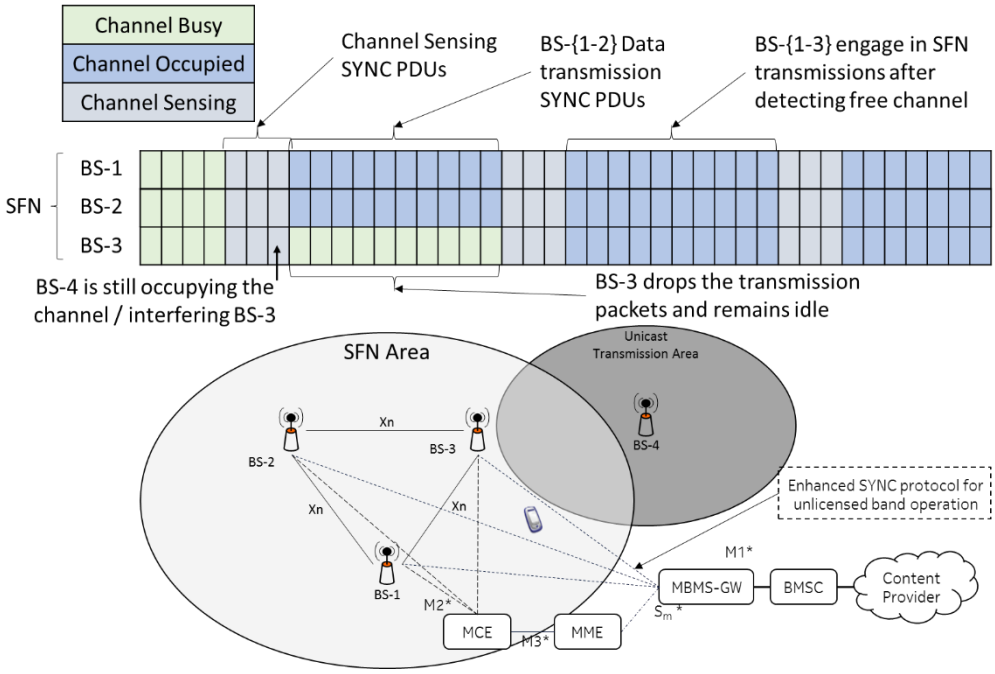


Figure A-12: Configurations for enabling D2D-augmented broadcast [76]

Adapting the mechanism to the 5G-Xcast system architecture shown in Sec. A.1 would require the enhanced synchronization mechanism for coordinating the transmission occasions to be located at the XUF / XCF for terrestrial broadcast transmissions, and within the radio access network for mixed-mode multicast transmissions. The functionality could be also located within the RAN for the terrestrial broadcast transmissions, assuming that the coordination for setting up the transmission area is done within the RAN. The practical use cases for the scenario could also be limited to localized rather than wide-area transmissions.

References

[1] 5G-Xcast Deliverable D4.1 v1.1: “Mobile Core Network”, June 2018.

[2] 5G-Xcast Deliverable D2.1 v1.1: "Definition of Use Cases, Requirements and KPIs", October 2017.

[3] 5G-Xcast Interim Deliverable D3.2 v1.0: "Air Interface", May 2018.

[4] 5G-Xcast Deliverable D2.2 v1.2: "Analysis of the Technical Developments Against the Use Cases", May 2018.

[5] ITU-T Recommendation I.112 (03/93): “Vocabulary of Terms for ISDNs”, March 1993.

[6] 3GPP TR 21.905 V15.1.0 (2018-12): “Vocabulary for 3GPP Specifications (Release 15)”, December 2018.

[7] 3GPP TS 23.501 V15.4.0 (2018-12): “System Architecture for the 5G System; Stage 2 (Release 15)”, December 2018.

[8] 3GPP TS 25.446 V15.0.0 (2018-07): “MBMS synchronisation protocol (SYNC) (Release 15)”, July 2018.

[9] 3GPP TS 36.300 V15.4.0 (2018-12): “Evolved Universal Terrestrial Radio Access (E-UTRA) and Evolved Universal Terrestrial Radio Access Network (E-UTRAN); Overall description; Stage 2 (Release 15)”, December 2018.

[10] 3GPP TS 36.321 V15.4.0 (2018-12): “Evolved Universal Terrestrial Radio Access (E-UTRA); Medium Access Control (MAC) protocol specification (Release 15)”, December 2018.

[11] 3GPP TS 38.300 V15.4.0 (2018-12): “NR; NR and NG-RAN Overall Description; Stage 2 (Release 15)”, December 2018.

[12] 3GPP TS 38.321 V15.4.0 (2018-12): “NR; Medium Access Control (MAC) protocol specification (Release 15)”, December 2018.

[13] 3GPP TS 38.322 V15.4.0 (2018-12): “NR; Radio Link Control (RLC) protocol specification (Release 15)”, December 2018.

[14] 3GPP TS 38.323 V15.4.0 (2018-12): “NR; Packet Data Convergence Protocol (PDCP) specification (Release 15)”, December 2018.

[15] 3GPP TS 37.324 V15.1.0 (2018-09): “E-UTRA and NR; Service Data Adaptation Protocol (SDAP) specification (Release 15)”, September 2018.

[16] 3GPP TS 38.331 V15.4.0 (2018-12): “NR; Radio Resource Control (RRC) protocol specification (Release 15)”, December 2018.

[17] 3GPP TS 38.401 V15.4.0 (2018-12): “NG-RAN; Architecture description (Release 15)”, December 2018.

[18] 3GPP TS 38.410 V15.2.0 (2018-12): “NG-RAN; NG general aspect and principles (Release 15)”, December 2018.

[19] 3GPP TS 38.411 V15.0.0 (2018-06): “NG-RAN; NG layer 1 (Release 15)”, June 2018.

[20] 3GPP TS 38.412 V15.1.0 (2018-09): “NG-RAN; NG signalling transport (Release 15)”, September 2018.

[21] 3GPP TS 38.413 V15.2.0 (2018-12): “NG-RAN; NG Application Protocol (NGAP) (Release 15)”, December 2018.

[22] 3GPP TS 38.414 V15.1.0 (2018-09): “NG-RAN; NG Data Transport; (Release 15)”, September 2018.

[23] 3GPP TS 38.420 V15.2.0 (2018-12): “NG-RAN; Xn general aspects and principles (Release 15)”, December 2018.

[24] 3GPP TS 38.421 V15.0.0 (2018-06): “NG-RAN; Xn layer 1 (Release 15)”, June 2018.

[25] 3GPP TS 38.422 V15.2.0 (2018-12): “NG-RAN; Xn signalling transport (Release 15)”, December 2018.
[26] 3GPP TS 38.423 V15.2.0 (2018-12): “NG-RAN; Xn application protocol (XnAP) (Release 15)”, December 2018.
[27] 3GPP TS 38.424 V15.1.0 (2018-09): “NG-RAN; Xn data transport (Release 15)”, September 2018.
[28] 3GPP TS 38.460 V15.2.0 (2018-12): “NG-RAN; E1 general aspects and principles (Release 15)”, December 2018.
[29] 3GPP TS 38.461 V15.0.0 (2018-06): “NG-RAN; E1 layer 1 (Release 15)”, June 2018.
[30] 3GPP TS 38.462 V15.2.0 (2018-12): “NG-RAN; E1 signalling transport (Release 15)”, December 2018.
[31] 3GPP TS 38.463 V15.2.0 (2018-12): “NG-RAN; E1 application protocol (E1AP) (Release 15)”, December 2018.
[32] 3GPP TS 38.470 V15.4.0 (2018-12): “NG-RAN; F1 general aspects and principles (Release 15)”, December 2018.
[33] 3GPP TS 38.471 V15.0.0 (2017-12): “NG-RAN; F1 layer 1 (Release 15)”, December 2017.
[34] 3GPP TS 38.472 V15.2.0 (2018-09): “NG-RAN; F1 signalling transport (Release 15)”, September 2018.
[35] 3GPP TS 38.473 V15.4.1 (2019-01): “NG-RAN; F1 application protocol (F1AP) (Release 15)”, January 2019.
[36] 3GPP TS 38.474 V15.2.0 (2018-09): “NG-RAN; F1 data transport (Release 15)”, September 2018.
[37] 3GPP Work Item Description RP-181349: “New SID Proposal: Study on Integrated Access and Backhaul for NR”, June 2018.
[38] 3GPP TR 38.874 V16.0.0 (2018-12): “NR; Study on Integrated Access and Backhaul; (Release 16)”, December 2018.
[39] 3GPP TS 29.281 V15.5.0 (2018-12): “General Packet Radio System (GPRS) Tunnelling Protocol User Plane (GTPv1-U) (Release 15)”, December 2018.
[40] IETF STD 6 / IETF RFC 768: “User Datagram Protocol”, August 1980.
[41] IETF STD 86 / IETF RFC 8200: “Internet Protocol, Version 6 (IPv6) Specification”, July 2017.
[42] IETF STD 5 / IETF RFC 791: “Internet Protocol”, September 1981.
[43] IETF RFC 4960: “Stream Control Transmission Protocol”, September 2007.
[44] 3GPP TR 36.890 v13.0.0, “Study on Support of single-cell point-to-multipoint transmission in LTE,” July 2015.
[45] 3GPP TR 38.913 V15.0.0: “Study on scenarios and requirements for next generation access technologies” (Release 14) July 2018.
[46] Fasil B. Tesema, Ahmad Awada, Ingo Viering, Meryem Simsek, and Gerhard P. Fettweis “Multiconnectivity for Mobility Robustness in Standalone 5G Ultra Dense Networks with Intra-frequency Cloud Radio Access”.
[47] 5G PPP mmMAGIC white paper, “Architectural aspects of mm-wave radio access integration with 5G ecosystem”, Version 1.0, April 2016.
[48] 5G-Xcast Deliverable D4.2: “Converged Core Network”, Version 2.0, October 2018.
[49] 3GPP TR 23.799 V14.0.0 (2016-12): “Study on Architecture for Next Generation System (Release 14)”, December 2016.
[50] 3GPP TR 38.801 V14.0.0 (2017-03): “Study on new radio access technology: Radio access architecture and interfaces (Release 14)”, March 2017.

[51] 3GPP TS 28.530 V15.1.0 (2018-12): "Telecommunication management; Management and orchestration; Concepts, use cases and requirements (Release 15)", December 2018.
[52] 3GPP TS 28.531 V16.0.0 (2018-12): "Management and orchestration; Provisioning; (Release 15)", December 2018.
[53] 3GPP TS 28.532 V15.1.0 (2018-12): "Management and orchestration; Generic management services; (Release 15)", December 2018.
[54] 3GPP TS 28.533 V15.1.0 (2018-12): "Management and orchestration; Architecture framework (Release 15)". December 2018.
[55] 3GPP TS 22.261 V16.6.0 (2018-12): "Service requirements for the 5G system; Stage 1 (Release 16)", December 2018.
[56] Emmanouil Pateromichelakis, Jens Gebert, Tomasz Mach, Jakob Belschner, Wei Guo, Nandish P. Kuruvatti, Venkatkumar Venkatasubramanian, and Caner Kilinc, "Service-Tailored User-Plane Design Framework and Architecture Considerations in 5G Radio Access Networks", IEEE Access, August 2017.
[57] Next Generation Mobile Networks (NGMN) Alliance, "NGMN 5G White Paper", Version 1.0, February 2015.
[58] 3GPP TR 22.852 V13.1.0 (2014-09): "Study on Radio Access Network (RAN) sharing enhancements (Release 13)", September 2014.
[59] 3GPP TR 22.951 V15.0.0 (2018-06): "Service aspects and requirements for network sharing (Release 15)", June 2018.
[60] Next Generation Mobile Networks (NGMN) Alliance, "Description of Network Slicing Concept", Version 1.0, January 2016.
[61] 3GPP TR 32.500 V15.0.0, "Telecommunication Management; Self-Organizing Networks (SON); Concepts and requirements", June 2018.
[62] 3GPP TS 32.501 V15.0.0, "Telecommunication management; Self-configuration of network elements; Concepts and requirements (Release 15)", June 2018
[63] 3GPP TR 36.902 V9.3.1, "Self-configuring and self-optimizing network (SON) use cases and solutions (Release 9)", March 2011.
[64] S. Hämäläinen, H. Sanneck, and C. Sartori, LTE Self-Organizing Networks (SON); Network Management Automation for Operational Efficiency. John Wiley & Sons, Ltd, 2012.
[65] S.Feng, E. Seidel, "Self-Organizing networks (SON) in 3GPP Long Term Evolution", Nomor White Paper, May 2008.
[66] 3GPP TR 36.742 V15.0.0 (2017-06): "Study on Self-Organizing Networks (SON) for enhanced Coordinated Multi-Point (eCoMP); (Release 15)", September 2018.
[67] SP-170658, "New WID on Self-Organizing Networks (SON) for Active Antenna System (AAS) deployment management", September 2017
[68] 3GPP TR 32.865 V15.0.0, "Telecommunication management; Study on Operations, Administration and Maintenance (OAM) aspects of Self-Organizing Network (SON) for Active Antenna System (AAS) based deployments; (Release 15)", June 2017
[69] SP-180827, "Study on Self-Organizing Networks (SON) for 5G", September 2018.
[70] 3GPP TS 26.346 v15.1.0, "Multimedia Broadcast Multicast Services (MBMS); Protocols and Codecs (Release 15)", June 2018.
[71] IETF RFC 6726, "FLUTE, File Delivery over Unidirectional Transport", November 2012 https://tools.ietf.org/html/rfc6726

[72] IETF RFC 5775, “Asynchronous Layered Coding (ALC) Protocol Instantiation”, April 2010, https://tools.ietf.org/html/rfc5775
[73] IETF RFC 6363, “Forward Error Correction (FEC) Framework”, October 2011, https://tools.ietf.org/html/rfc6363
[74] IETF RFC 4588, “RTP Retransmission Payload Format”, October 2011, https://tools.ietf.org/html/rfc4588
[75] A. Prasad, M.A. Uusitalo, D. Navaratil, M. Säily, ”Challenges for Enabling Virtual Reality Broadcast Using 5G Small Cell Networks,” IEEE WCNC Workshops (WCNCW), April 2018.
[76] A. Prasad, A. Maeder, M.A. Uusitalo, “Optimized Over-The-Air Virtual Reality Broadcast Transmissions with Low-Latency Feedback,” IEEE 5G World Forum (5G-WF), July 2018.
[77] A. Prasad, P. Lunden, M.A. Uusitalo, Z. Li, ”Enhancements for Enabling Point-to-Multipoint Communications Using Unlicensed Spectrum,” IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB), June 2018.