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

Deliverable D4.1
Mobile Core Network

Version v1.0 Date: 2018/05/31

Document properties:

Grant Number:	761498
Document Number:	D4.1
Document Title:	Mobile Core Network
Editor(s):	Tuan Tran (EXP)
Authors:	Carlos Barjau Estevan, David Gomez-Barquero (UPV); David Navratil, Athul Prasad (NOK); Jon Hart (BT); Maël Boutin (Broadpeak); Roman Odarchenko, Rui Andrade Aguiar (Bundeslab); Christophe Burdinat, Tuan Tran (Expway); Khishigbayar Dushchuluun, Christian Menzel (IRT); Baruch Altman (Live U); Peter Sanders, Menno Bot (one2many); Castagno Mauro (TIM)
Contractual Date of Delivery:	2018/05/31
Dissemination level:	Public
Status:	Final
Version:	v1.0
File Name:	5G-Xcast_D4.1_v1.0

Abstract

This document describes the 5G mobile core network that enables multicast and broadcast capabilities where two different alternatives have been considered. This document provides an analysis of the architectural alternatives. The architectural alternatives have been built based on the 5G-Xcast design principles and building blocks. Two different approaches to leverage multicast and broadcast capabilities in 5G mobile core network have been studied: transparent multicast transport and point-to-multipoint services.

This document also presents new functionalities and technologies considered within 5G-Xcast such as converged autonomous switch between unicast, multicast and broadcast for the converged network including fixed broadband and mobile networks, Public Warning for multimedia content, Multilink and Multi-access Edge Computing. In addition, the limitations of eMBMS in LTE that have motivated the current work are also outlined in this document.

Keywords

5G, 5GC, architecture, broadcast, design principle, MBMS, converged middleware, Multi-access edge computing (MEC), mobile network, MooD, multicast, Multilink, point-to-multipoint, Public Warning (PW)

Executive Summary

This document describes the 5G-Xcast mobile core network architecture that enables multicast and broadcast capabilities based on 5G architecture defined in 3GPP Release 15. The proposed architectures will be the target for the technical developments within other Work Packages (WP) of the project such as WP3 on the interaction between Radio Access Network (RAN) and the core network and WP5 on the content distribution framework. The 5G-Xcast mobile core network architecture considers the new functionalities and technologies such as converged autonomous switch between unicast, multicast and broadcast for the converged network including fixed broadband and mobile networks, multimedia public warning alert, multi-connectivity and multilink, and multi-access edge computing.

The use cases and requirements identified in Work Package 2 (WP2) in the 5G-Xcast project are reviewed from the core network point of view. These technical requirements are considered as the basis for providing the 5G-Xcast architecture solutions.

This document describes the 5G-Xcast design principles that enables multicast and broadcast capabilities. These design principles are aligned with the ones identified in 3GPP Release 15.

Focusing on multicast and broadcast capabilities in the mobile core network, this document provides two different approaches to the same problem of network resource optimization: the transparent multicast transport and point-to-multipoint services including multicast, broadcast and terrestrial broadcast.

Most importantly, this document describes two primary architecture alternatives to enable multicast and broadcast capabilities inside 5G core network architecture:

- The first alternative leverages the 5G core network architecture to enable multicast and broadcast capabilities.
- The second alternative aims at minimizing the changes to the LTE eMBMS specification while being compatible with the 5G core network architecture.

This document also takes into consideration another possible alternative which does not introduce any new network functions to the 5G core network architecture. However, this alternative could be considered as an implementation option of the first alternative.

Table of Contents

List of Figures

List of Tables

List of Acronyms and Abbreviations

5GC	5G Core Network
ABR	Adaptive Bitrate
AF	Application Function
AL-FEC	Application Layer Forward Error Correction
AMF	Access and Mobility Management Function
API	Application Programming Interface
AR/VR	Augmented Reality / Virtual Reality
AS	Application Server
ATSSS	Access Traffic Steering, Switching and Splitting
AUSF	Authentication Server Function
AVP	Attribute Value Pair
BM-SC	Broadcast Multicast Service Centre
BPM	Broadcast Provisioning Manager
CAPEX	Capital Expenditure
CBC	Cell Broadcast Centre
CBE	Cell Broadcast Entity
CDN	Content Delivery Network
CGI	Cell Global Identifier
CMAS	Commercial Mobile Alert System
CR	Consumption Report
CUPS	Control and User Plane Separation
DASH	Dynamic Adaptive Streaming over HTTP
DC	Dual Connectivity
DCI	Downlink Control Information
DRA	Diameter Routing Agent
DN	Data Network
DRB	Data Radio Bearer
DRM	Digital Right Management
DSL	Digital Subscriber Line
DTLS	Datagram Transport Layer Security
DTT	Digital Terrestrial Television
DVB	Digital Video Broadcasting
EC	European Commission
ECGI	E-UTRAN Cell Global Identifier
eMBMS	Evolved Multimedia Broadcast Multicast Services
FEC	Forward Error Correction
GCS	Group Communication System
GERAN	GSM EDGE Radio Access Network
GSM	Global System for Mobile Communications
GTP	GPRS Tunnelling Protocol
HLS	HTTP Live Streaming
HTHP	High Tower High Power
HTTP	Hypertext Transfer Protocol
IE	Information Element
IGMP	Internet Group Management Protocol
IoT	Internet of Things
ISD	Inter-Site Distance
ISI	Inter-Symbol Interference
JSON	JavaScript Object Notation
KPI	Key Performance Indicators
LTE	Long-Term Evolution

mABR	Multicast adaptive bit rate
MBMS	Multimedia Broadcast Multicast Service
MBMS-GW	MBMS Gateway
MBMS SAI	MBMS Service Area Identity
MBSFN	Multicast Broadcast Single Frequency Network
MCCH	Multicast Control Channel
MCData	Mission Critical Data
MCE	Multi-cell/multicast Coordination Entity
MCPTT	Mission-critical push-to-talk
MEC	Multi-access Edge Computing
ML	Multilink
MLD	Multicast Listener Discovery
MME	Mobility Management Entity
MNO	Mobile Network Operator
MooD	MBMS operation on Demand
MPD	Media Presentation Description
MP-TCP	Multipath TCP
MTCH	Multicast Traffic Channel
MW	Middleware
NAS	Non-Access Stratum
NEF	Network Exposure Function
NF	Network Function
NRF	NF Repository Function
NSSF	Network Slice Selection Function
OAM	Operations, administration and maintenance
OFDM	Orthogonal frequency-division multiplexing
PCF	Policy Control Function
PDCP	Packet Data Convergence Protocol
PDN	Packet Data Network
PDSCH	Physical Downlink Shared Channel
PDU	Protocol Data Unit
P-GW	PDN Gateway
PIM	Protocol Independent Multicast
PSS	Packet-switched Streaming Service
PTM	Point to Multipoint
PW	Public Warning
QoE	Quality of Experience
QoS	Quality of Service
QUIC	Quick UDP Internet Connections
RAB	Radio Access Bearer
RAN	Radio Access Network
ROM	Receive Only Mode
RRC	Radio Resource Control
SAI	Service Area Identifier
SC-MRB	Single Cell MBMS Point to Multipoint Radio Bearer
SC-PTM	Single Cell Point to Multipoint
SCTP	Stream Control Transmission Protocol
S-GW	Serving Gateway
SIB	System Information Block
SINR	Signal-to-Interference-plus-Noise Ratio
SMF	Session Management Function
TDF	Traffic Detection Function
TMGI	Temporary Mobile Group Identity
UDM	Unified Data Management

UDR	Unified Data Repository
UDSF	Unstructured Data Storage network function
UE	User Equipment
UPF	User Plane Function
V2X	Vehicle to Everything
WLAN	Wireless Local Area Network
XCF	5G-Xcast Control Plane Network Function
XUF	5G-Xcast User Plane Network Function

Terminology

Throughout the project 5G-XCast the usage of the terms “broadcast” and “multicast” has special importance. In particular, the term “broadcast” is understood in different ways by broadcasters and by mobile industry.

For historical reasons broadcasters understand by “broadcast” systems that possess only a downlink to distribute their content in a Point-to-Area mode (as for e.g. DAB or DVB-T). Usually these systems are operated in spectrum bands allocated to “broadcast service” and that do not provide an uplink. In 3GPP, “Receive-Only-Mode (ROM) broadcast” was first introduced with 3GPP Release 14 to provide an equivalent service. For “ROM broadcast”, the UE need not register/attach with/to the network. Before 3GPP Release 14 “ROM-broadcast” was not supported by 3GPP.

The usage of the term “broadcast” within the mobile industry originates from mobile systems that are always operated in spectrum allocated to mobile services, i.e. that have both uplink and downlink, and the UE is registered/attached to/with the network. This type of “broadcast” together with “multicast” has first been specified by 3GPP in Release 6 within MBMS. eMBMS for LTE was introduced in 3GPP Release 9 and supported only this type of “broadcast”, which is a point-to-multipoint content delivery method for UE that is required by the specification to register/attach with/to the network for the “unicast” operation. That means, the UE is always capable of “unicast” communication with the network, although UE’s “unicast” communication capability may not be required for the point-to-multipoint content delivery method in cases when the associated procedures (for e.g. the file repair or the reception reporting) are not used. In this case of “broadcast” the UEs do not need to join the delivery session as with “multicast”.

For the sake of clarity about the terminologies, the term “terrestrial broadcast” used in the present document refers to the term “ROM broadcast”, i.e. to “broadcast” as used by the broadcasters. The terms “broadcast” and “multicast” are used in the present document according to the above explained understanding by 3GPP. In addition, the term “point-to-multipoint” in the present document refers to all “multicast”, “broadcast” and “terrestrial broadcast”. Further definitions about the terminologies are described in the deliverable D2.1 “Definition of Use Cases, Requirements and KPIs” [2]. It’s important to note that “Xcast” as part of 5G-Xcast project name is not related to the term Xcast described in [50].

5G-Xcast aims at proposing the solutions for all “multicast”, “broadcast” and “terrestrial broadcast” type of transmissions.

1 Introduction

The first 3GPP release of 5G technology (Release 15), also known as 5G NR (New Radio), will be completed in 2018. Release 15 has been structured in three phases. An early drop at the beginning of 2018 of a non-standalone (NSA) version that requires LTE for the control plane. The 5G NR NSA leverages not only the LTE ePC (evolved Packet Core), but also the LTE RAN (Radio Access Network) for wide coverage and mobility. It will introduce 5G NR to enhance the user plane performance and efficiency using dual connectivity across the LTE and NR bands. During the second-half of 2018, the stand-alone (SA) version of 5G will be standardized, including the 5G core network (5GC), that will enable deployments without any LTE infrastructure. The 5G NR SA deployment can be also in combination with LTE but using only 5G NR for the user plane as in the early drop. The last drop of Release 15 specification is expected at the end of 2018, and it will enable more architecture options for hybrid LTE and 5G NR deployments using the 5GC. It will basically enable using the 5GC to inter-work with both LTE RAN and NR RAN, using the NR RAN for the control plane.

An impending problem of the first release of 5G is that it only supports unicast communications in the core network and point-to-point (PTP) transmissions in the RAN. This limitation may imply an inefficient service provisioning, and utilization of the network and spectrum resources when distributing the same data to multiple users and devices. One of the 3GPP system requirements for the 5G system is to provide flexible multicast/broadcast services [1], since it is considered as an essential feature for 5G applications in a number of vertical sectors. Deliverable D2.1 of 5G-Xcast [2] describes use cases for point-to-multipoint (PTM) transmissions in 5G for four different verticals: media and entertainment, automotive (V2X communications), Internet-of-Things IoT (machine-type communications), and Public Warning & Safety (critical communications). Another vertical that requires PTM transmissions is Airborne Communications (e.g. drone communications).

5G multicast/broadcast is one of the topics that is under discussion at 3GPP for 5G phase II (Release 16). Two different tracks have been identified: (i) “Terrestrial Broadcast” and (ii) “Mixed Mode Multicasting” [3]. “Terrestrial Broadcast” enables a dedicated downlink-only broadcast-only network suitable for Digital Terrestrial Television (DTT) in both High-Tower High-Power (HTHP) and Low-Tower Low-Power (LTLP) deployments leveraging cellular technology. For the “Terrestrial Broadcast” track, 3GPP will use the LTE Release 14 EnTV (Enhancements Television Services) work [4] as a basis and enhance it if needed to meet the 5G requirements for multimedia broadcast services, [1] (clause 6.13) and [5] (clause 9.1). A gap analysis will be performed to evaluate which requirements are not met by LTE Release 14. “Mixed Mode Multicasting” allows for dynamic mode switching between unicast PTP and multicast PTM to more efficiently deliver identical content. “Mixed Mode Multicasting” implies both multicast and broadcast used throughout the present document. For the “Mixed Mode Multicasting” track, 3GPP will use 5G NR as a basis, and it has been acknowledged that diverse use cases from media, IoT, V2X and public safety should be considered when designing the solution.

5G-Xcast is very well aligned to the “Mixed Mode Multicasting” track in 3GPP, since the project vision is to incorporate point-to-multipoint capabilities in 5G as built-in delivery features for network optimization, integrating point-to-point and point-to-multipoint modes under one common framework and enabling dynamic use of point-to-multipoint to maximize network and spectrum efficiency [6], [7]. Furthermore, 5G-Xcast will also

design a “Terrestrial Broadcast” mode based on 5G NR (RAN and core network), based on the mixed mode multicasting solution.

This deliverable presents the 5G-Xcast mobile core network architecture that introduces multicast and broadcast capabilities in the 5GC. The 5G-Xcast mobile core network architecture has been built upon the 3GPP 5G network architecture while its first release (Release 15) has been, and it is currently still, under standardisation. The 5G-Xcast mobile core network is very well aligned with 3GPP key principles and concepts for the 5G core network, such as service-based interface between control plane network functions (NF), function separation and modularization, and the CUPS (control and user plane separation) design principle. Two different architecture solutions are proposed in this deliverable. One solution provides an approach that is 5G architecture friendly. The other solution provides minimal changes to the eMBMS architecture and specification described for LTE. This deliverable also takes into consideration another possible alternative that does not introduce any new NFs to the 5GC but it’s considered as an implementation option of the first solution. Both solutions aim to integrate multicast/broadcast communication capabilities as a built-in optimization feature of the core network. While it is generally understood that the proposed core network architectures can support terrestrial broadcast, the applicability of the proposed core architectures for terrestrial broadcast will be studied further in the second half of the project.

The rest of the document is structured as follows. Section 2 provides a high-level description of eMBMS architecture as background on how multicast and broadcast communications are currently supported by a 3GPP System. Section 3 provides a brief overview of the current 3GPP 5G core network architecture (Release 15). Section 4 describes the new functionalities and technologies that 5G-Xcast project aims to enable in the 5G mobile core network. Section 5 presents the design principles of the 5G-Xcast core network and Section 6 describes the 5G-Xcast building blocks and network functions. Section 7 describes the proposed 5G-Xcast core network architectures and their analysis. Finally, Section 8 concludes the document and discusses the open topics that will be addressed in the second half of the project. Annex A includes an analysis of the limitations of 3GPP eMBMS Release 14 from a core network perspective. Recall that the analysis of the limitations of 3GPP eMBMS Release 14 from a RAN perspective can be found in Deliverable D3.1 [8]. Hence, the terminologies in this section are used based on the view from 3GPP. Annex B describes the localized eMBMS architecture for V2X in Release 14 to minimize the latency.

2 eMBMS Architecture in LTE

The following high-level description of eMBMS (Evolved Multimedia Broadcast Multicast Services) architecture provides a background on how multicast and broadcast communications are currently supported by a 3GPP system. It is based on 3GPP Release 14 since Release 15 3GPP 5G / new radio (NR) system does not currently support any type of multicast and/or broadcast communications.

For the sake of clarity, it is anticipated the support of multicast/broadcast communications (by the 5G system) in a *4G-like* fashion would not be seen as adequate from a 5G-Xcast perspective since the analysis of (Release 14) 3GPP eMBMS identified the following limitations, regarding:

- Lack of dynamic configuration for MBSFN
- Lack of feedback on eMBMS session management
- Lack of feedback from RAN to the core network
- Lack of efficient mechanism to trigger MBMS reception in the UE
- Service continuity when moving between MBSFN areas

Further details on the 3GPP eMBMS architecture limitations from a 5G-Xcast perspective are provided in the Annex A.

2.1 Overview of eMBMS architecture

eMBMS is able to provide multicast and broadcast multimedia services through the LTE network, combining unicast with MBMS data in the same LTE radio frame as specified in 3GPP TS 36.300 [9] and TS 23.246 [10] . eMBMS introduced in 3GPP Release 9 re-uses the features of 3G MBMS specified from Release 6. The eMBMS architecture is shown in Figure 1 with the following entities:

- **BM-SC** (Broadcast Multicast Service Centre) is located at the core network. The BM-SC provides functions for MBMS User Service provisioning and delivery to the content provider. It can also serve as an entry point for MBMS data traffic from the MBMS User Services.
- **MBMS-GW** (MBMS Gateway) is a logical entity that serves as an entry point for IP multicast traffic as MBMS data from the BM-SC. The MBMS-GW uses IP multicast as the means of forwarding MBMS user data to the eNodeB(s). The MBMS-GW sends MBMS Session Control Signalling (Session start/update/stop) towards the downstream node (i.e. MME) which routes the signalling messages to MCE serving the broadcast area.
- **MCE** (Multi-cell/multicast Coordination Entity) provides the admission control and the allocation of the radio resources used by all eNodeBs in the MBSFN area for multi-cell MBMS transmissions using MBSFN operation. As part of radio resource allocation procedure, the MCE decides the radio configuration such as modulation and coding scheme for both control and data. In addition, the MCE also decides on whether to use SC-PTM or MBSFN. SC-PTM and MBSFN delivery modes are described in section 2.1.2.

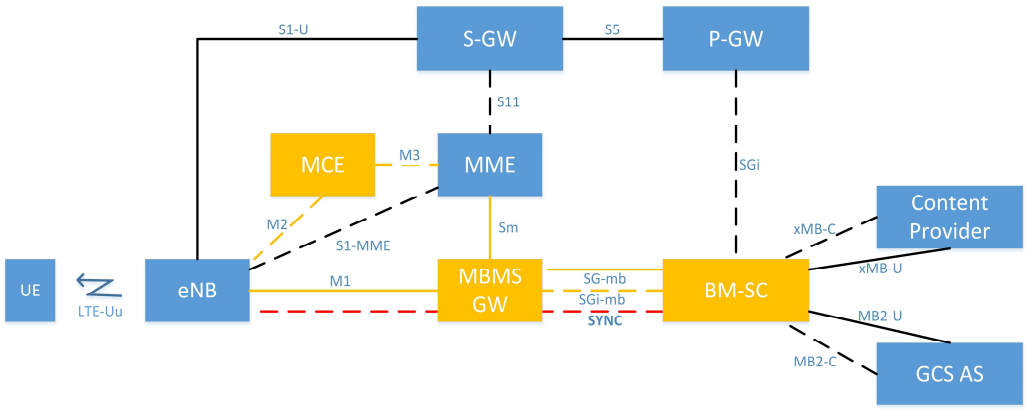


Figure 1: eMBMS Architecture.

The BM-SC comprises of the following functions:

- Membership Function: This function provides authorization for UEs requesting to activate an MBMS service. This function may have subscription data of MBMS service users or generate charging records for MBMS service users. This function exists only in case of multicast mode (see clause 4.4.1 of 3GPP TS 23.246 [10]) but not for broadcast mode (see clause 4.4.3 of 3GPP TS 23.246 [10]).
- Session and Transmission Function: This function controls the Session Start/Stop/Modification procedures via SGmb interface. It also allocates TMGI (Temporary Mobile Group Identity) to MBMS User Service.
- Proxy and Transport Function: This is a Proxy Agent for signalling over SGmb reference point between MBMS-GWs and other BM-SC sub-functions, e.g. the BM-SC Membership Function and the BM-SC Session and Transmission Function. It can be separated into additional two sub-functions: Proxy for managing the control plane and Transport for managing multicast payload.
- Service Announcement Function: This function which is a user service level function provides the UE with media descriptions or MBMS session descriptions specifying the media or the MBMS sessions to be delivered as part of an MBMS user service.
- Security Function: This function provides integrity and/or confidentiality protection for MBMS Data. This function is used for distributing MBMS keys to authorized UEs.
- Content synchronization for MBMS: This sub-function applies the SYNC protocol [11] to the MBMS Data.

There are two alternatives to deploy an MCE as shown in Figure 2. In the centralized MCE architecture (left part of Figure 2), the MCE is a logical entity which means it can be deployed as a stand-alone physical entity or collocated in another physical entity (e.g. eNodeB). In both cases of the centralized MCE architecture, the M2 interface is kept between the MCE and all eNodeB(s) controlled by this MCE. In the distributed MCE architecture (right part of Figure 2), an MCE is part of the eNodeB and the M2 interface is internal to the eNodeB.

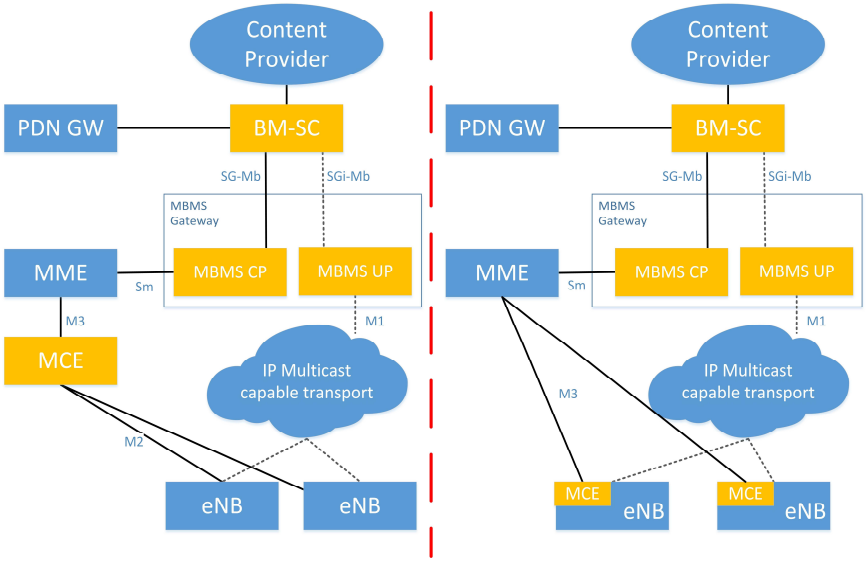


Figure 2: Centralized versus Distributed MCE deployment.

2.1.1 Interfaces and reference points related to eMBMS architecture

Figure 1 shows the eMBMS architecture with interfaces and reference points between the entities.

xMB

xMB is the reference point between Content Provider and BM-SC and is specified in 3GPP TS 26.346 [12] and TS 29.116 [13]. This interface has been introduced in 3GPP Release 14. The interface is based on JSON over HTTP. Using this reference point, a content provider can invoke procedures supported by BM-SC(s) to setup and manage an MBMS user service from BM-SC to the MBMS clients.

MB2

The MB2 interface introduced in 3GPP Release 13 is specified in 3GPP TS 23.468 [14] for stage 2 and in 3GPP TS 29.468 [15] for stage 3. This interface provides session management between the GCS AS and BM-SC for MCPTT services. MB2 carries control plane signalling (via MB2-C) and user plane data (via MB2-U) between the GCS AS and BM-SC. The data transferred via MBMS bearer(s) is delivered from the BM-SC using the Group Communication delivery method as defined in 3GPP TS 26.346 [12].

SGmb and SGi-mb

The SGmb interface uses diameter protocol and the interface is specified in 3GPP TS 29.061 [16] as an IETF vendor specific diameter application. The interface provides session management between the BM-SC and MBMS-GW. SGi-mb is the reference point between BM-SC and MBMS-GW for MBMS data delivery. The SGi-mb interface is used by BM-SC to send user plane data over UDP/IP. MBMS-GW uses the destination IP address and the destination UDP port to determine to which MBMS bearer (M1 interface) to forward the received data.

Sm

The Sm interface uses GTPv2-C protocol as specified in 3GPP TS 29.274 [17]. This interface provides session management between the MBMS-GW and the MME.

M1

M1 is the user plane interface between an eNodeB and MBMS-GW for MBMS data delivery. The M1 interface uses GTP-U protocol over UDP and IPv4 or IPv6 multicast. The payload of GTP-U are SYNC packets. The M1 interface structure is specified in 3GPP TS 36.445 [18].

M2

M2 is the interface between MCE and eNodeB which allows for MBMS session control signalling, radio configuration data and radio access counting. The interface implements M2 application protocol as specified in 3GPP TS 36.443 [19] which runs over SCTP (Stream Control Transmission Protocol).

M3

M3 is the interface between MME and MCE which allows for MBMS session control signalling at the e-RAB level and uses M3 application protocol specified in 3GPP TS 36.444 [20]. This interface does not convey radio configuration data.

2.1.2 MBSFN and SC-PTM delivery modes

eMBMS originally had only MBSFN (Multicast Broadcast Single Frequency Network) as broadcast delivery mode. Although the support of single cell broadcast was discussed already in 3GPP Release 8 (partially included and later removed from Release 8 specification), it was only in 3GPP Release 13 when single cell broadcast in form of SC-PTM (Single-Cell Point To Multipoint) was introduced as a second mechanism of MBMS delivery in addition to MBSFN. The main motivation is to meet the latency and coverage granularity requirements of Mission Critical Services.

2.1.2.1 MBSFN

MBSFN transmission is a simulcast transmission technique where multiple cells transmit identical waveforms at the same time and the same frequency. An MBSFN transmission from multiple cells within the MBSFN Area can be considered as a single transmission by a UE. Thanks to the use of OFDM (orthogonal frequency-division multiplexing) and Cyclic Prefix (CP) at the physical layer, signals received inside the Cyclic Prefix contribute positively while signals received outside this interval would create interference between OFDM Symbols otherwise called ISI (Inter-Symbol Interference).

The Cyclic Prefix is a subgroup of the last samples of the OFDM symbol that is transmitted at the beginning of the OFDM symbol. Hence, Cyclic Prefix leads to an overhead and lowers the capacity. Cyclic Prefix duration is a trade-off between overhead and robustness against multipath interference.

SFN networks provide a gain in coverage, especially at the cell edge inside the SFN area due to the positive signal contribution from multiple cells. SFN networks improve the signal-to-interference-plus-noise ratio (SINR) on the condition that the received signals have a propagation delay less than the Cyclic Prefix duration. The improved SINR and the fact that transmission is received by multiple UEs are the main contributors which make eMBMS transmissions more efficient than unicast transmission, even for a relatively small number of users receiving in the same content.

According to 3GPP TS 36.300 [9], several concepts regarding the eMBMS broadcast provision area are defined. Figure 3 shows the relationship between these concepts:

- MBSFN Area consists of a group of cells within an MBSFN Synchronization Area, these cells are coordinated to achieve an MBSFN transmission. MBSFN Synchronization Area is an area of the network where all eNodeBs can be synchronized and can be configured to perform MBSFN transmission. Reserved cells are used to reduce the interference. Reserved cells are cells within a

MBSFN Area that do not contribute to the MBSFN transmission. Normally, reserved cells are located at the edge of a MBSFN area, usually forming a ring as shown in Figure 4, where no MBSFN signals are transmitted (i.e. no (P)MCH carrying MCCH and MTCH is broadcasted).

- MBMS Service Area is the geographic area where the mobile network operator (MNO) decides to deliver a MBMS service. This region consists of one or more MBSFN Areas. MBMS Service Areas comprises of cells assigned with one or several MBMS SAI (Service Area Identifier). Each MBMS SAI addresses one or more cells. Since one cell can belong to more than one eMBMS service, one cell can be associated with several MBMS SAIs. MBMS Service Area can be served by one or more MBSFN areas.

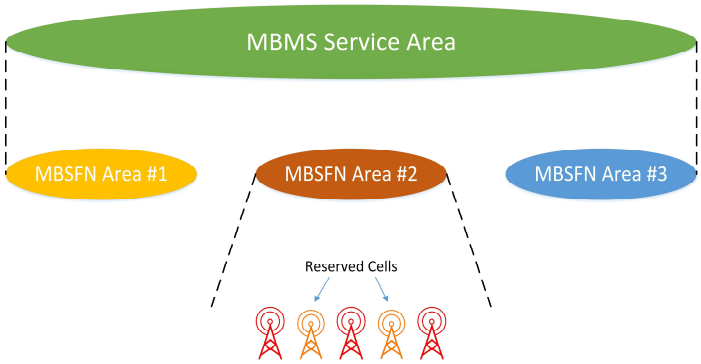


Figure 3: MBMS definitions.

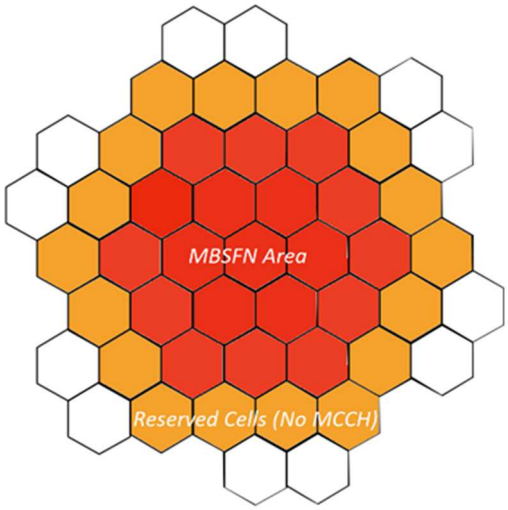


Figure 4: Example of MBSFN network with a ring of Reserved Cells.

2.1.2.2 SC-PTM

SC-PTM provides an efficient delivery of content for the scenarios where the UEs are located in small geographical area at the coverage granularity of a few cells while MBSFN transmission provides the data to all cells in MBSFN area. SC-PTM reuses the eMBMS architecture with the same logical entities and interfaces, as shown in Figure 1.

In SC-PTM, the UEs receive the content through a common radio resource region in PDSCH (Physical Downlink Shared Channel). This concept allows the multiplexing between normal unicast data and broadcast data within the same PDSCH subframe.

Further differences between SC-PTM and MBSFN from the radio perspective are described in the deliverable D3.1 “LTE-Advanced Pro Broadcast Radio Access Network Benchmark” [8].

3 Current 3GPP 5G Core Network Architecture

The system architecture for 5G is defined in 3GPP TS 23.501 [21]. The procedures and Network Function Services for the 5G architecture are defined in 3GPP TS 23.502 [22]. Both specifications are developed in the Release 15 timeframe. The 5G architecture is defined as service-based and the interaction between network functions (NF) is represented in two ways:

- A service-based representation, where network functions within the control plane enables other authorized network functions to access their services. This representation also includes point-to-point reference points where necessary.
- A reference point representation, shows the interaction that exists between the network function services described by point-to-point reference point between any two network functions.

Figure 5 depicts the non-roaming service-based reference architecture. Service-based interfaces are used within the control plane.

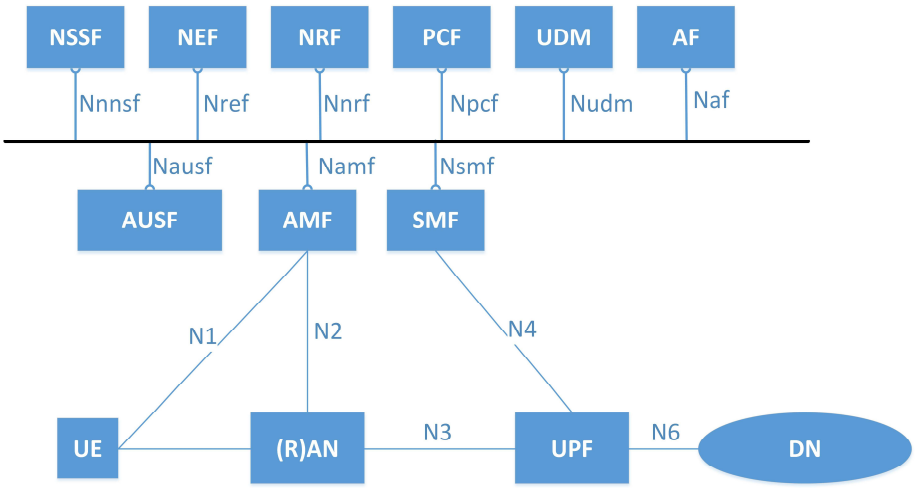


Figure 5. 5G System architecture in service-based representation.

Figure 6 depicts the 5G System architecture in the non-roaming case, using the reference point representation showing how various network functions interact with each other.

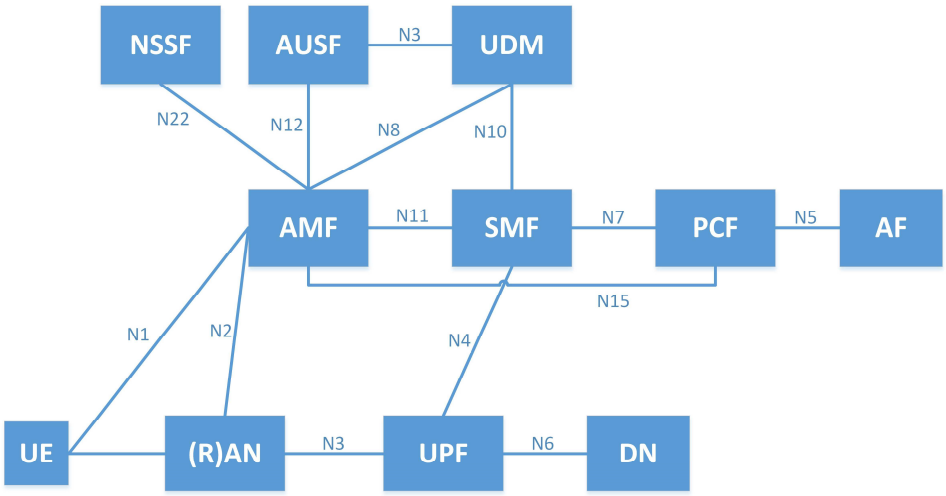


Figure 6. Non-Roaming 5G System Architecture in reference point representation.

The network functions and entities are specified in clause 4.2.2 of 3GPP TS 23.501 [21]. It's noted that not all NFs are shown in Figure 6. The 5G System architecture consists of the following network functions:

- Authentication Server Function (AUSF) for the authentication of UEs when they register to the network.
- Access and Mobility Management Function (AMF) for access management and mobility management of UEs.
- Data Network (DN), e.g. operator services, Internet access or 3rd party services.
- Unstructured Data Storage network function (UDSF) supports storage and retrieval of information as unstructured data by any network function.
- Network Exposure Function (NEF) to securely expose network functions provided by the 3GPP network to for example 3rd parties, Application Functions and Edge Computing.
- NF Repository Function (NRF) supports the service discovery function to allow Network Functions to discover other Network Functions.
- Network Slice Selection Function (NSSF) allows selecting the set of network slice instances serving the UE and determining the AMF set to be used to serve the UE.
- Policy Control Function (PCF) for the support of a unified policy framework to govern network behaviour.
- Session Management Function (SMF) for management of UE sessions and policy enforcement.
- Unified Data Management (UDM) stores subscription data and authentication data.
- Unified Data Repository (UDR) supports storage and retrieval of subscription data by the UDM and policy data by the PCF.
- User Plane Function (UPF) for user data packet routing and forwarding between UE and DN.
- Application Function (AF) is an instantiation for the control plane part of an application, such as an MBMS application.
- User Equipment (UE).
- (Radio) Access Network ((R)AN)

Based on operator deployment, an Application Function that is considered to be trusted can be allowed to interact directly with relevant network functions and is therefore not restricted to interact with only the PCF and the NEF. In case the AF is not allowed by the operator to access directly the NFs, the AF uses the NEF to interact with relevant NFs.

MBMS is not supported yet in the 5G architecture because 3GPP has not prioritized MBMS for Release 15. As per 5G-Xcast design principles, multicast and/or broadcast are rather considered as a tool for internal optimization, thus the MBMS functions are considered to be a trusted application service in 5G-Xcast.

4 5G-Xcast New Functionalities and Technologies

This section describes the new functionalities and technologies that the 5G-Xcast project aims to enable in 5G. Section 4.1 first describes the current 3GPP MooD which enables a seamless experience whenever the network switches between unicast and MBMS delivery to the UEs only in mobile network. Section 4.1 also presents a new concept of converged autonomous 5G-Xcast multicast operation on demand ("5G-Xcast MooD"), which is applied in a converged network including both fixed and mobile networks. In addition, 5G-Xcast MooD applies for both transparent multicast transport and point-to-multipoint services as described in section 6. Section 4.2 presents the current Public Warning (PW) functionalities based on Cell Broadcast technology and explains why the current technology is not able to support PW multimedia data. Section 4.3 presents the multilink concept to improve the available bandwidth, reliability, mobility, etc. by using simultaneously multiple available connections. Section 4.4 describes Multi-access Edge Computing (MEC) to enable new applications and services in 5G (e.g. AR/VR).

4.1 5G-Xcast MooD

3GPP MooD (MBMS operation on Demand) enables the dynamic establishment of MBMS User Services according to actual consumption, in order to offload unicast content delivery and to efficiently use network resources when the traffic volume exceeds a certain threshold. 3GPP MooD in mobile networks (for e.g. LTE) is standardized in 3GPP Release 12 specified in 3GPP TS 26.346 [12] and TR 26.849 [23]. A high-level architecture is shown in Figure 7.

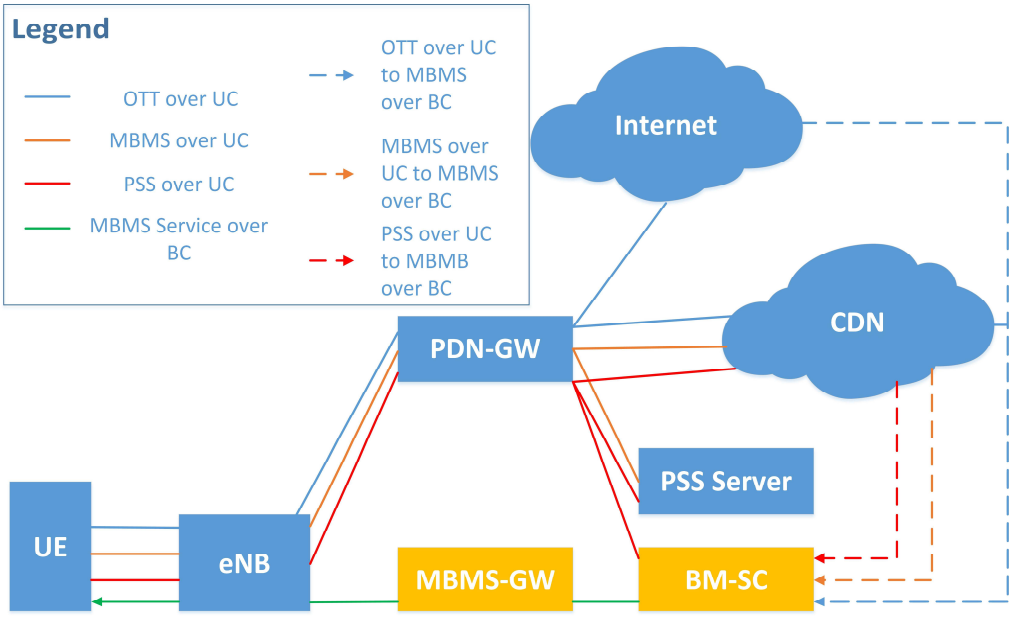


Figure 7. High-level 3GPP MooD architecture.

Upon the BM-SC's determination of high attachment rate of a unicast service, it activates a MBMS user service to carry the same content over the MBMS bearer. This conversion is illustrated by the unicast service whose transport is shown by the blue, red or orange coloured line to an MBMS user service carried on the MBMS bearer as shown by the green-coloured line. The 3GPP specification (Figure 7) does not show explicitly the building blocks and/or the application logics to enable 3GPP MooD. Figure 8 shows the 3GPP MooD solution architecture, especially for DASH content.

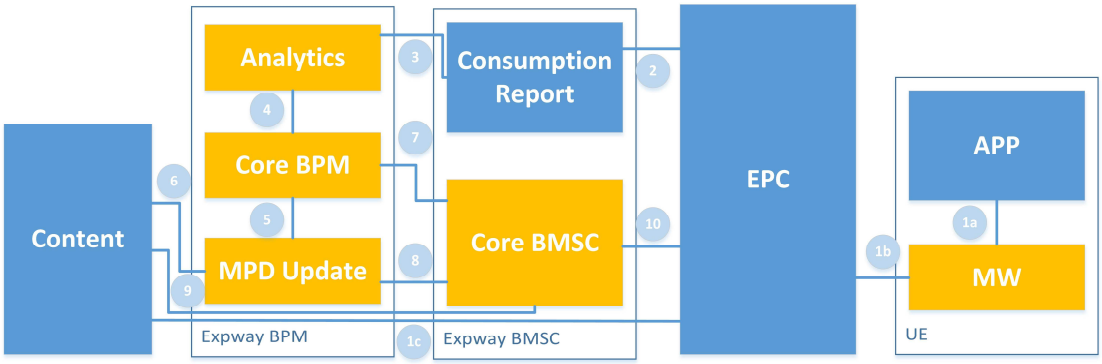


Figure 8. 3GPP MooD solution architecture.

1) The application requests the content to the middleware (MW) (also known as MBMS Client as defined in 3GPP TS 26.347 [24]. The MW acts as a proxy server. If the content is delivered in unicast, the MW receives directly the content from the content source (1c).

2) The MW periodically sends the Consumption Reports to the Consumption Report Server through the EPC.

3) The Analytics Server pulls the information contained in Consumption Report or the Consumption Report Server pushes the information to the Analytics Server. The Analytics Server then performs the statistics about the content consumption.

4) The Analytics Server decides to deliver a popular content over MBMS.

5/6) The Broadcast Provisioning Manager (BPM) updates the manifest (MPD for DASH content) in order to add multicast and unicast profiles (Unified MPD).

7) The BPM instructs the BM-SC to deliver the content over MBMS.

8/9) The BM-SC retrieves the unified MPD and video segments.

10) The BM-SC performs Service Announcement and delivers content to UEs.

Consumption Reports (CR) is used for audience measurement. The CR mechanism is similar to MBMS Reception Reports specified in 3GPP TS 26.346 section 9.4 [12]. The CR contains multiple attributes such as *samplePercentage*, *reportInterval*, *clientId*, *locationType*, etc. The UE's location type is used to decide when to trigger unicast/multicast switching. Possible values of the *locationType* attribute are "CGI" (Cell Global Identifier), "ECGI" (E-UTRAN Cell Global Identifier) or "MBMS SAI" (MBMS Service Area Identity).

3GPP MooD was specified before SC-PTM (3GPP Release 13). From the standard perspective, 3GPP MooD is available for MBSFN delivery mode. However, a cell-based MBMS offloading (e.g. SC-PTM) seems to be possible if the *locationType* attribute is set to ECGI.

In the fixed broadband network, there isn't any standard multicast on demand solution although proprietary solutions exist. 5G-Xcast will enable 5G-Xcast MooD where the seamless unicast/multicast/broadcast switching can be performed in mobile or fixed networks. Furthermore, the seamless switching can also be performed when the UE switches from fixed to mobile network and vice versa. The reader is referred to the deliverable D5.2 on "Key Technologies for Content Distribution Network" (to be published in September 2018) [25] which describes the aspirations in this area. 5G-Xcast MooD is not a single solution but a collection of tools, procedures and best practices applied in various parts of end-to-end system as shown in the deliverable D5.2 to enable efficient

multicast delivery. 5G-Xcast MooD may comprise two stages. In the first stage, a content is provided in a format suitable for multicast including transport, i.e. IP multicast. In the second stage, the (radio) access network optimizes its resources by selecting the best transport mode possible. For example, a radio access network such as 5G RAN determines a set of point-to-point and point-to-multipoint bearers to deliver the content to the receiving UEs in the most efficient way.

4.2 Public Warning

Public Warning (PW) in LTE is specified in the Cell Broadcast Service specification (3GPP TS 23.041 [26]). Cell Broadcast allows sending limited size (text) messages to UEs in a specific area. Figure 9 shows the PW architecture.

Figure 9: PW architecture.

A Cell Broadcast Entity (CBE) is an entity on which PW messages are composed by a message originator and submitted to the Cell Broadcast Centre (CBC). There is currently no standardized protocol for this interface. However, ATIS has specified a protocol in the ATIS-070008 standard [27], which offers downlink-only capabilities for message broadcasting.

The CBC determines the Cell IDs and the Tracking Area IDs for those cells, that cover the target area indicated by the CBE. The CBC formats the message according to the SBc protocol as specified in 3GPP TS 29.168 [28]. Such a message contains, amongst others, a List of Tracking Areas element which is used by the MME to determine the eNodeBs that need to be addressed. The CBC is unaware of eNodeBs and the MME is aware which eNodeBs serve the Tracking Areas in the List of Tracking Areas. The message also contains a Warning Area which contains a Cell ID List, a Tracking Area ID list or an Emergency Area ID list. The eNodeB uses the Warning Area element to determine which cells need to broadcast the message.

It should be noted that the MME has no knowledge of cells that are controlled by an eNodeB; it knows only Tracking Areas that are served by the eNodeB. It should be also noted that Cell ID, Tracking Area ID and Emergency Area ID are the IDs which are configured in the eNodeB for each cell that it serves. In practice the area is hardly ever used.

The MME forwards the message, without a List of Tracking Areas, to the eNodeBs as per S1-MME protocol defined in 3GPP TS 36.413 [29]. It is noted that the MME is part of the PW architecture although mobility plays no role in the PW function. It was a decision of 3GPP SA WG2 to include the MME as a proxy between the CBC and the eNodeBs.

The eNodeB receives the message with the Warning Area element which contains all the cells that are being addressed in the entire target area, even though the eNodeB itself cannot support more than 256 cells. The MME applies no filtering on the Warning Area list. The eNodeB finds from the Warning Area element the cells that it supports and that need to transmit the message.

The eNodeB first transmits a paging message which includes a CMAS-indication (Commercial Mobile Alert System) and is repeated every 40 ms. When the UE receives a paging message which includes this CMAS-indication then the UE will acquire SIB1 (System Information Block) message which will contain the scheduling information for

SIB12, which is repeated every 80 ms. SIB12 contains the warning message content and is typically repeated every 320 ms (the range is 80 ms – 5,12 s).

Depending on the air interface bandwidth and the configured DCI format and the size of the text message, a SIB12 message may have to be segmented. The UE will have to receive all segments in order to extract the message content and if one segment is missed, it will have to wait until this segment is repeated.

The message in SIB12 contains, apart from the text itself, a Serial Number, a Message Identifier and a Data Coding Scheme parameter. The Serial Number and Message Identifier combination is used to determine duplicate messages, since Cell Broadcast messages are repeated and shall not be displayed more than once. The Message Identifier indicates the source or topic of the message, such as an Amber alert or an imminent threat message. The Data Coding Scheme parameter indicates how the message was encoded and possibly may contain a language indication which can be used for filtering.

The maximum size of a text message is 1230 octets, consisting of a maximum of 15 CB pages of 82 octets each. The formatting of text messages into pages goes back to GSM: the GERAN air interface can only broadcast a single page per message. A multi-page message requires the message to be segmented into multiple messages for broadcasting. Access technologies after GSM allow messages with more payload, but the formatting of text messages has remained access technology independent. The maximum size of a binary message is 9600 octets. The value of 9600 octets is to be considered unrealistic given the issues with segmentation. Smartphones these days support PW text messages, but generally do not support binary messages.

It's obvious that 1230 octets or even 9600 octets are not enough to carry multimedia messages. The work done in this project will seek to deliver PW multimedia messages using multicast/broadcast capability defined in this project.

4.3 Multi-connectivity and Multilink

4.3.1 Multi-connectivity technologies

Multi-connectivity (MC) of single user terminal to multiple radio access points is a 5G key enabler in order to satisfy the demanding requirements on 5G mobile networks. Multi-connectivity supports simultaneous connectivity and aggregation across different technologies such as 5G, LTE, and unlicensed technologies such as IEEE 802.11 (Wi-Fi) (Figure 10). In addition, a single user terminal may connect to multiple network layers such as macro and small cells and multiple radio access technology (RAT) layers such as below 6GHz and mmWave. In heterogeneous networks, multi-connectivity helps to provide an optimal user experience (for e.g. high bandwidth, network coverage, reliable mobility).

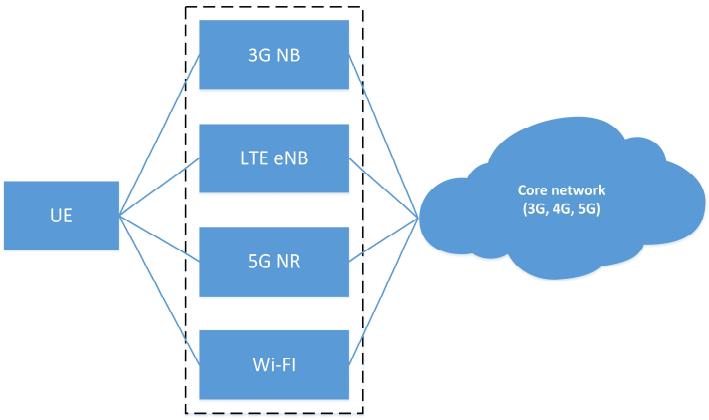


Figure 10. 5G architecture with Multi-connectivity.

As mentioned above MC can be provided at different network layers (e.g. micro/macro), spectrum (e.g. sub-6 GHz/mm-wave), user plane (e.g. PDCP (Packet Data Convergence Protocol)), technologies (e.g. Wi-Fi/LTE) depending on service, deployment and RAT. In this document, we present two specific architecture options which allow for integrating multi-connectivity into 5G networks (Figure 11)

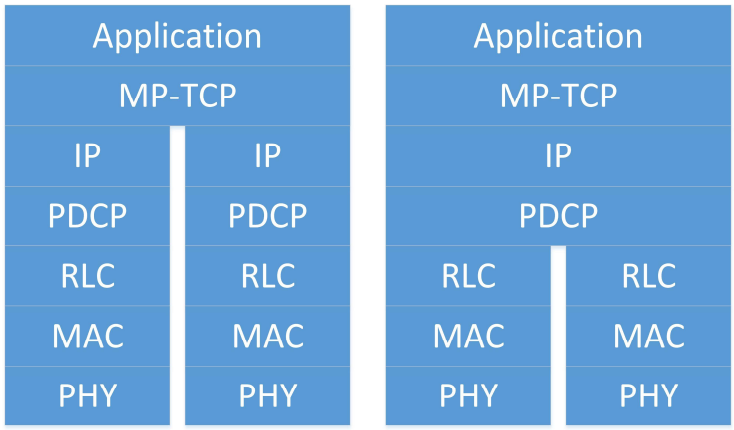


Figure 11. Options for integrating multi-connectivity into 5G networks.

The user plane aggregation among multiple novel 5G radio technologies could take place on PDCP level (or their 5G equivalents). PDCP-level aggregation can enable several features with the benefits of being likely more suitable for distributed deployments with non-ideal backhaul and not requiring the harmonization of the lower layers of the radio technologies.

This section describes some well-known MC technologies, such as Dual connectivity (DC), LTE-WLAN Aggregation (LWA), RAN-Controlled LTE-WLAN Interworking (RCLWI), LTE-WLAN Radio Level Integration with IPsec Tunnel (LWIP).

Proposed in 3GPP Release 12, Dual Connectivity (DC) is a capability of a device to connect to two eNodeBs (of the same technology) simultaneously (Figure 12). The base stations are connected via X2 interface, hence enabling direct flow of packets through split bearer. The dual connectivity approach enhances reliability of data flow. However, it does not address the scenario if the two base station belongs to different RATs.

Multi-RAT Dual Connectivity (MR-DC) is a generalization of the Intra-E-UTRA Dual Connectivity (DC) described in 36.300 [9], where a multiple Rx/Tx UE may be configured to utilise radio resources provided by two distinct schedulers in two different nodes connected via non-ideal backhaul, one providing E-UTRA access and the other one providing NR access. One scheduler is located in the master node (MN) and the other in the secondary node (SN). The MN and SN are connected via a network interface and at least the MN is connected to the core network.

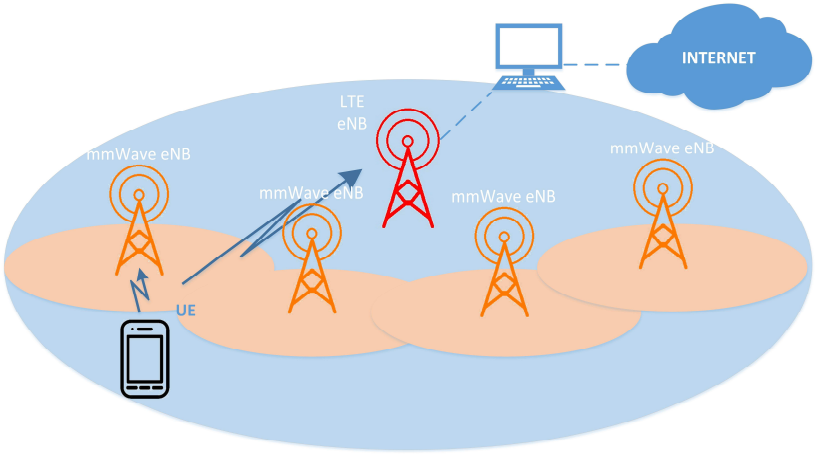


Figure 12. Dual-connectivity example.

In some cases, parallel usage of available Wi-Fi and cellular networks can also provide better user experience. The following mechanisms have been standardized within 3GPP Release 13 under the LTE-Wi-Fi RAN-level integration framework:

- **LTE-WLAN Aggregation (LWA)**: is an evolution of Dual Connectivity, where the secondary link is provided by the Wi-Fi Access Point (AP). This mechanism involves very tight resource aggregation, where a single DRB can be either switched very fast between LTE and Wi-Fi link or split and provided simultaneously by the two RATs. However, in order to be able to do that, the Wi-Fi network needs to be upgraded with the WT (WLAN Termination) logical entity and support Xw interface. Additionally, the UE needs to be upgraded with LTE-WLAN Aggregation Adaptation Protocol (LWAAP), to be able to properly route the PDCP PDUs coming from Wi-Fi link.

- **RAN-Controlled LTE-WLAN Interworking (RCLWI)**: is also based on WT and Xw interface upgrade of the Wi-Fi network for control signalling, however, the User Plane (UP) bearers instead of going through the LTE eNodeB are routed through a CN with Wi-Fi legacy link. This is rather a bearer handover (or an offload) than an aggregation compared to LWA, however still the UE is controlled by the network to receive the data from Wi-Fi link, instead of taking this decision by itself. Compared to LWA, this solution doesn't require the UE to be upgraded with LWAAP.
- **LTE-WLAN Radio Level Integration with IPsec Tunnel (LWIP)**: provides the possibility to aggregate resources from Wi-Fi and LTE simultaneously (similar to LWA), but without requiring the upgrade of the Wi-Fi network (i.e. enabling the use of legacy Wi-Fi networks). The Wi-Fi link is managed by the LTE eNodeB, however instead of the LWA-like flow control and use of LWAAP, an IPsec tunnel is established between UE and eNodeB. The splitting of bearer is not possible as the aggregation is done at IP level.

As described above, Release 13 has brought different options for very tight network-controlled LTE-Wi-Fi integration. This will enable different levels of integration between Wi-Fi and LTE; and depending on the required Wi-Fi network upgrade and/or UE side upgrade:

- LWA provides the tightest resource aggregation, whereas it requires highest level of upgrade on the different entities involved.
- LWIP allows the use of legacy network, still enabling the resource aggregation on the RAN level.
- RCLWI on the other hand requires an upgrade at the network side but doesn't require UE to be upgraded. This mechanism doesn't allow very tight resource aggregation as the Wi-Fi link is anchored at the CN side.

4.3.2 Multilink

In this document, we will use the term multilink (ML) to refer to various combinations of IP links aggregation. For example:

1. 5G cellular network link of one operator with a Wi-Fi network link of the same or another operator.
2. 5G cellular network link of one operator with a cable, xDSL or satellite IP link of same or another operator.
3. Any number of IP links of same or different operators

In the scenarios that require high bandwidth or assured service continuity, a user may need multiple concurrent connections (see Figure 12) For example, data aggregation from multiple subscriptions to LTE, 3G and Wi-Fi (and even fixed networks) increases available bandwidth. A cellular (e.g. 5G or LTE) network access is required to maintain the service continuity after a UE has access to Wi-Fi coverage. In these strategies, a ML-GW (Multilink[1] Gateway) is able to reroute the data packets through the different available links, and a ML-MW (Multilink middleware) performs the adequate data merger operation at the UE. The ML-MW at the viewing user side communicates with the ML-GW which can be located either at the core network, the publisher, or the cloud depending, for example, on deployment constrains. These two entities (ML-GW and ML-MW) exchange information about the performance of each link.

The content transmitted from the ML-GW down to the viewing device is split or duplicated over available links which are possibly from different operators or uses different technologies or according to their temporal performance. The decision whether to split or to duplicate depends on the desirable gains in throughput, ancillary information and reliability, and a function of the link conditions. The content is then reassembled at the viewing device (with eventual duplicates removed) as a coherent data stream ready for viewing. The content itself is not manipulated which means that the delivery is completely agnostic to the content.

[1] The specific usage of terminology will be cleared later in this section.

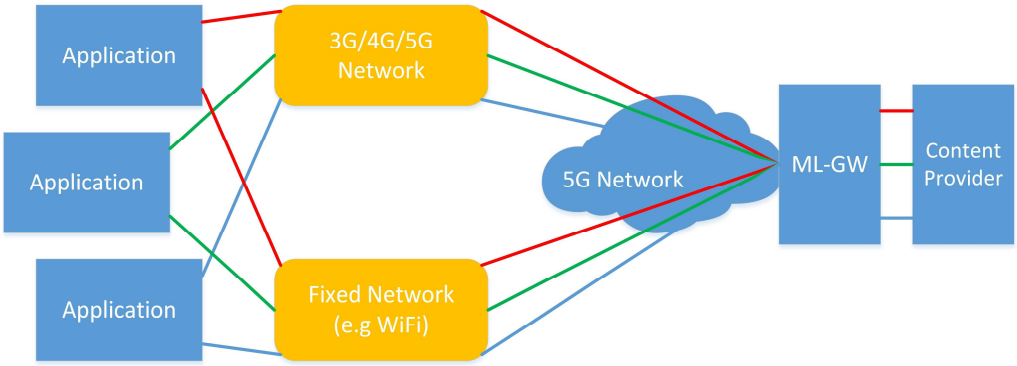


Figure 13. 5G Architecture with Multilink.

Figure 13 presents multi-connectivity from the perspective of heterogeneous network connections (and from the IP protocol side, the concept is known as multihomed devices).

Note that the specific relation between IP routing and ML is a research point that needs to be addressed inside the global architecture and may result in multi-layered solutions. The end-to-end packet delivery may (or may not) be encapsulated and (virtual) tunnels built, or multihomed IP may be used. Overall control plane design will need to be carefully considered as well as the relation between the SMF (Session Management Function) and the UPF (User Plane Function).

4.3.3 Multi-connectivity usage

There are several options for implementing ML strategies, which are also related to the key objective of the ML usage. ML can be used for improving reliability, providing ancillary information, increasing available bandwidth (e.g. higher video resolution), or performing traffic optimization. Due to this diversity, the use of the multiple connections can be quite distinct in different models. The main methods for multi-connectivity can be described as:

1. “Replicate”: Deliver the same content over all available connections.
2. “Switching”: Send all the content on one link and then switch between links as a fail-over mechanism.
3. “Load balancing”: Balance the content between all links so that each data flow, or IP destination, or application, use one of the links.
4. “Complementing”: A specific load balancing approach, which uses the key stream information on the primary link and provide additional information on the secondary link.
5. “Bonding” or “Aggregation”: Treat all available links as a single virtual broadband link and split all the content between all available links dynamically according to the performance of each of the links and the total available throughput.

It is noted that not all of these methods use multiple connections simultaneously. In the 5G-Xcast project, we shall mostly pursue the last one, which are the most powerful and challenging. In particular, the benefits of bonding can be listed as follows:

1. Overall bandwidth: The possibility to deliver broadband content that would be impossible to deliver over a single link. For example, if a certain video stream needs 15 Mb/s and the single link is capable to deliver only 10 Mb/s, then this content could not pass over a single link. However, when bonding and aggregating at the application layer at least two such links, the total available bandwidth becomes 20 Mb/s which makes this delivery possible. This technology could be advantageous to 5G since video content (especially with current high

video resolution trend) is usually either a live stream or a very large file to download, and individual links even 5G ones, are not always sufficient.
2. Reliability and availability of the service: In any single layer-2 link, especially in a wireless environment, the fluctuation in bandwidth, latency or error rate can be dramatic. Using multiple links as a virtual single broadband connection could mitigate these fluctuations. Seamless transition between single-L2-link and multilink could be achieved in a reliable way due to the use of simultaneous multiple networks in a dynamic way.
3. Mobility support: the first two feature benefits also imply that the mobility (including at high speed) is supported in an improved manner. In this case, mobility means the seamless transition between coverage areas of different networks or technologies, with continuous QoS and QoE. For instance, the end user can enjoy a seamless viewing experience when moving from the office to the home using the same mobile device.

The main steps in the bonding at the IP layer are:

1. Evaluate in real time the changes in application-level performance of each link (e.g. “goodput”, latency, jittery behaviour) in each of the relevant directions (e.g. uplink, downlink).
2. Evaluate the total available “goodput” at each point in time.
3. Split the ongoing data stream to all available links according to the performance of each link (i.e. not “overload” any of these links). Note that there are scenarios where some content may not be delivered to all terminals.
4. Buffer is necessary at the UE side to accommodate out-of-order packet arrival, missing packets etc.
5. Combine the split content from multiple links into one common stream.
6. In some cases (e.g. live video), it is possible to add an integrated video encoding process which outputs a just-in-time transcoded video stream that adaptively matches the momentary performance of the multiplicity of virtually bonded links.

Multi-link aggregation is currently implemented only for unicast streams, bringing together distinct unicast connections to support a stream. The performance of each available link is measured between ML-GW and ML-MW by exchanging the information. Therefore, simple (in the sense that it has no feedback measurement mechanism) multicast or broadcasts over a common set of links lacking measurement information will not work in many cases. One of the challenges in 5G-Xcast project is to see how a sophisticated non-naive broadcast or multicast strategy over multi-link can be achieved. In this case, 5G-Xcast will benefit not only from seamless transitions between broadcast and multicast to unicast (and vice versa), and/or seamless unicast experience side by side with broadcasts to others, but also higher-quality broadcast will be enabled by using multiple connections. The other challenge of 5G-Xcast is related to the specific usage of multi-connectivity in wireless.

4.4 Multi-access Edge Computing

The multi-access edge computing (MEC) paradigm aims at exploring the potential that could be achieved through the convergence of diverse fields such as communication and information technology (IT). Such a convergence would lead to the development of new applications and services enabled by the provisioning of cloud computing at the edge of the fixed and/or wireless access network. The overall mobile edge computing framework presented in [30], adapted to a converged network is as shown in Figure 14. Currently envisioned key use cases for MEC from a 5G-Xcast perspective include video analytics, IoT, mass delivery of augmented/virtual reality (AR/VR), data caching and optimized local content distribution. MEC could play a key role in hosting the low-latency VR/AR applications which could then be delivered to the end user clients using fixed/mobile

access networks. The caching of frequently fetched content at the edge can allow MNOs to significantly optimize the transport network load, thereby minimizing deployment costs.

One possible application of MEC from a 5G-Xcast perspective, is as shown in Figure 15, where the VR application and related high-capacity, low-latency content is hosted. The mass delivery of high-volume data transmissions requires significant scaling of the transport network necessitating increased CAPEX investments for MNOs, which can be minimized if the content and related application is hosted in the edge of the access network. The low-latency constraints for VR traffic requires the content to be hosted closer to the access network, with possible dynamic update of the viewed content enabled through the low-latency application layer feedback between the VR application client in the end user device and the host server.

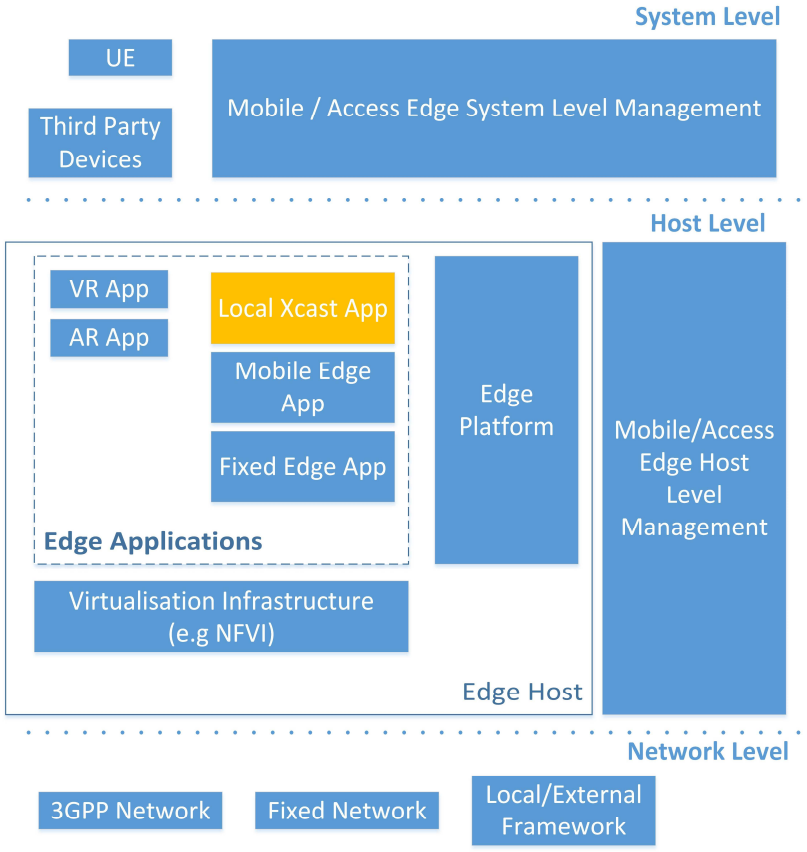


Figure 14: Mobile edge computing framework adapted to multi-access.

Edge computing can be supported by 5G architecture in many ways. The key enablers in 5G architecture for the support of edge computing are described in 3GPP TS 23.501. The 3GPP 5G system architecture defines a generic network function, the application function, for the interaction with 3GPP 5G core network (5GC). The application function may for example influence traffic routing, interact directly or indirectly via the network exposure function with the other network functions and use the policy framework for policy control.

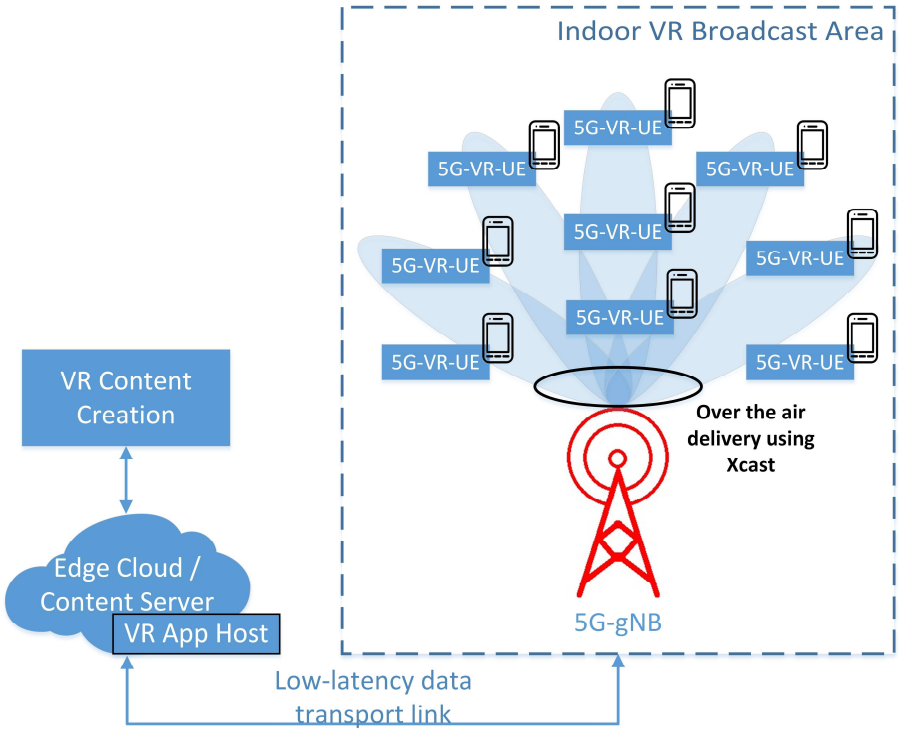


Figure 15: Possible application of MEC in 5G-Xcast.

MEC platform takes the place of the application function in the 5G system architecture as illustrated in Figure 16. The network functions may offer services to the MEC platform, which exposes the services to MEC application via the mobile edge platform application enablement framework including radio network information APIs and others [31], [32], [33], [34]. The MEC applications are located in the local data network. The 5G core network executes the traffic steering from the UPF to the local data network of MEC platform via a N6 interface towards MEC application.

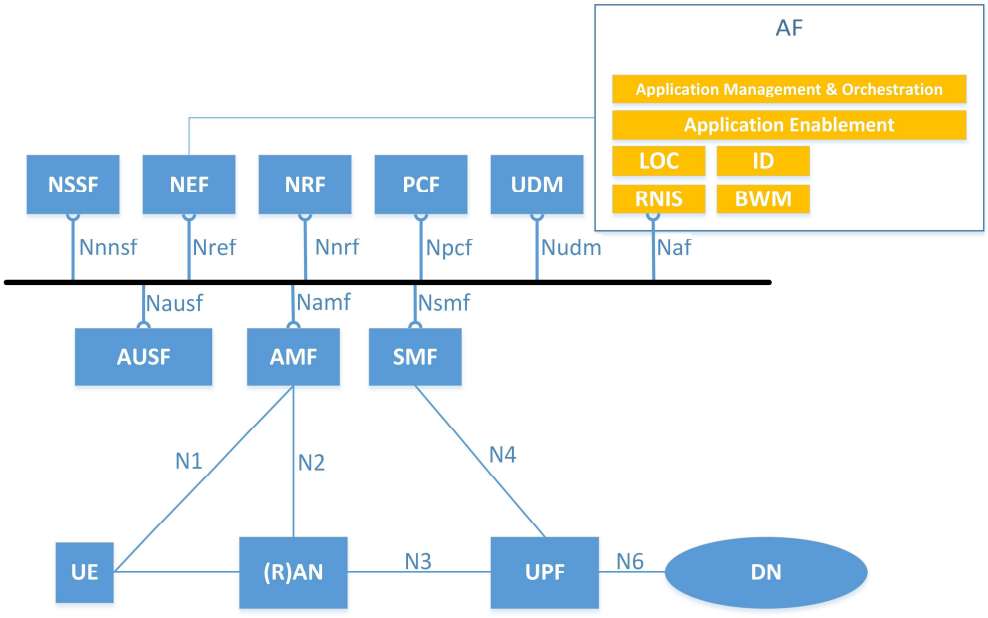


Figure 16: MEC in 5G system architecture.

5 5G-Xcast Mobile Network Design Principles

5.1 Design principles

5G-Xcast mobile network should build upon the 3GPP 5G network architecture whose first release (Release 15) is under standardisation. In 5G, 3GPP leverages service-based interaction between control plane network functions. Clause 4.1 of 3GPP TS 23.501 provides some key principles and concepts as follows:

- Separate the User Plane (UP) functions from the Control Plane (CP) functions, allowing independent scalability, evolution and flexible deployments, for e.g. centralized location or distributed (remote) location.
- Modularize the function design, for e.g. to enable flexible and efficient network slicing.
- Wherever applicable, define procedures (i.e. the set of interactions between network functions) as services, so that their re-use is possible.
- Enable each NF to interact with other NF directly if required. The architecture does not preclude the use of an intermediate function to help route Control Plane messages (e.g. like a DRA).
- Minimize dependencies between the Access Network (AN) and the Core Network (CN). The architecture is defined with a converged core network with a common AN - CN interface which integrates different Access Types, for e.g. 3GPP and non-3GPP access.
- Support a unified authentication framework.
- Support "stateless" NFs, where the "compute" resource is decoupled from the "storage" resource.
- Support capability exposure function.
- Support concurrent access to local and centralized services. To support low latency services and access to local data networks, UP functions can be deployed close to the Access Network.
- Support roaming with both Home routed traffic as well as Local breakout traffic in the visited PLMN.

The design principles in 5G-Xcast are aligned with 3GPP direction. This document provides additional principles related to multicast and broadcast capabilities from a core network perspective while other 5G-Xcast deliverables [25] provide their respective design principles such as from the Radio Access Network and Content Delivery Framework.

- Enabling multicast and broadcast capabilities should require a small footprint on top of the existing unicast architecture.
- Wherever possible, treat multicast and broadcast as an internal optimization tool inside the network operator’s domain.
- Consider terrestrial broadcast as a service offered by MNO for UE without uplink capabilities that can be delivered as a self-containing service by subset of functions of multicast and broadcast architecture.
- Simplify the system setup procedure to keep the system cost marginal. The design aims to develop an efficient system in terms of architecture/protocol simplicity and resource efficiency. Despite simplified procedures, the architecture also should allow flexible session management.
- Focus on the protocols that allows efficient IP multicast.
- Enable caching capabilities inside the network.

Although this document focuses on mobile core network, the network design is viewed from converged network perspective where the fixed, mobile and possibly broadcast networks are put in place.

5.2 Considerations for multicast and broadcast in 5G-Xcast mobile network architecture

3GPP has not considered multicast and broadcast capabilities in Release 15, thus, they are not included in 3GPP TS 23.501 v1.5.0 of 2017-11 [21]. Taking into consideration that the user plane and the control plane are separated, the following considerations apply to the 3GPP 5G architecture for multicast and broadcast capabilities.

- In the user plane, a NF could receive user data over a reference point, process that data (e.g. encapsulate, apply application forward error correction) and send the processed data to the RAN, where the data should be transmitted to UEs, possibly via other NFs. The NF receiving user data from the content provider could support the user plane of xMB protocol (or its evolution in 5G) for content ingestion.
- In the user plane, unicast user data and multicast or broadcast user data could pass through the same instances of user plane NFs (i.e. network resources are shared between unicast and multicast) or through different instances of user plane NFs allowing for functions such as traffic steering optimizations and load balancing.
- In the control plane, needed control functions could reside in an instantiation of an AF which could also provide an interface to a content provider such as the control plane of xMB protocol (or its evolution in 5G). It could be also possible to utilize and enhance the policy and charging framework for multicast and broadcast purposes.

It is assumed that the multi-cell multicast coordination function (MCE function in LTE) will reside in the RAN. It is also assumed that the SMF will be responsible for the corresponding session management. Network functions and overall architecture are discussed in detail in the following sections.

6 5G-Xcast Building Blocks and Network Functions

6.1 Considerations on the design principles and constraints

6.1.1 Modularization and function separation

Function separation and modularization are key design principles of 5G network which allow for a great flexibility in deployment and operation. The flexibility is needed because 5G networks ought to satisfy requirements of various use cases in different vertical domains. The network slicing concept uses the flexible network design to deliver customized networks to meet these requirements with minimum or ideally no overhead. The overhead can be understood as an unnecessary deployment of certain redundant functions, which are not needed to fulfil use case's requirements. For example, if we consider the following two cases:

A) A UPF implements all multicast/broadcast functionalities.
B) Multicast/broadcast functionalities are implemented by one or more separate NFs.

In case B, a network slice instance that does not require multicast/broadcast will not include the NF needed for multicast/broadcast. On the other hand, the UPF implemented to support multicast/broadcast in case A could be instantiated in a network slice instance, which does not require the multicast/broadcast functionalities. In case A, there may be an implementation and run-time overhead while in case B, the standardized architecture already tries to address some implementation optimization. The 3GPP specification in Release 15 does not require an instance UPF to support all UPF functionalities. This means that the overhead in case A is not understood as an issue from the specification point of view. However, a question arises whether all multicast/broadcast functionalities, which are needed to enable point-to-multipoint services, should be considered as a user plane functionality related to a PDU session or whether they are higher layer functionalities (i.e. a user plane seen from PDU session level) or control plane functionalities. In the example of a reliable file delivery that relies on a data storage and a repair procedure, the repair procedure may require a feedback from receiving UEs about the success or the failure of file delivery. The feedback can be a user plane functionality, a higher layer functionality, or a control plane functionality depending on a solution. The design process should consider based on this and other aspects of point-to-multipoint services if the case B approach is more appropriate. The network slicing for multicast/broadcast capabilities is discussed in the deliverable D4.2 "Converged Core Network" (to be published in September 2018) [25].

Sometimes, overhead is unavoidable, owing to design decisions involving modularization and functional separation, however in such cases the benefits of the good, modular design should outweigh the induced overhead.

An approach to network architecture design which aims at minimizing the overhead could design the network functions as fine-grained as possible. This approach has been adopted recently in service-oriented application architectures known as microservices where an application is a collection of loosely coupled microservices and lightweight protocols. The microservice architectures have proven to be successful in cloud applications. However, the same approach adopted for 5G network design could lead to a significantly large number of standardized interfaces. This introduces new design constraints in addition to the minimization of the overhead in deployments.

The network functions for 5G-Xcast should be designed so that efficient network slicing is possible with minimum overhead while keeping the number of network functions and thus the number of required standardized interfaces at a reasonable level.

5G-Xcast takes into account the following two approaches to enable point-to-multipoint (i.e. multicast, broadcast and terrestrial broadcast) capabilities:

- the first approach which is described in section 6.1.2 considers a transparent multicast transport inside the 5G network operator;
- the second approach which is described in section 6.1.3 addresses the point-to-multipoint capability as a service.

6.1.2 Transparent multicast transport

The transport of multicast traffic requires an elementary network function needed to deliver multicast data through the network to the user equipment (UE). For example, in the case of IP traffic, this means the transport of IP multicast datagrams. It is important that the 5G network supports this elementary function, which is enough to satisfy certain use cases such as LTE for mission critical application (known as Mission Critical Push-to-Talk (MCPTT) [35]). In MCPTT, eMBMS is used to transport data transparently to the UE [36]. It can be expected that the support of transparent transport of multicast user data will become an important function for future use cases such as IoT or IPTV in converged networks. For example, MCPTT relies on the MB2 interface for the transparent transport of user data via eMBMS [15]. The transparent transport of user data can be further simplified in 5G networks by allowing the 5G network to receive multicast user data directly, without a need for a tunnel from the multicast source when the multicast source resides inside the operator's 5G network, e.g. in multi-access edge cloud, or near the 5G network in operator's infrastructure and the networking between the multicast source and 5G network supports multicast transport.

Implementation-specific or standard-based solutions can be introduced to deliver multicast data to the UPF when the network infrastructure between the 5G network and the multicast source does not support multicast transport. The operator's infrastructure may be connected to the multicast content source by a virtual private network (VPN). This is considered as an implementation-specific solution. A standard-based solution could be a standardized API through which the multicast content source can request an establishment of a tunnel end-point, e.g. using DTLS. The transparent multicast transport means that the UPF treats the multicast data (e.g. IP multicast datagrams) in the same or very similar way as the unicast data even if multicast data are delivered to the UPF via a tunnel. For example, the UPF does not perform any multicast specific functionalities such as application layer forward error correction (AL-FEC) or file repair. A UE launches the file repair procedure when this UE cannot recover the missing data even after AL-FEC recovery. It is assumed that any protocols or optimization for unidirectional transport may implement the transparent multicast transport capability of the 5G system by the system (e.g. DVB multicast ABR) or the applications. In transparent multicast transport, specific functionalities (e.g. reliability from retransmission or AL-FEC) may be performed at the multicast content source rather than the 5G system providing such functionalities.

The transparent multicast transport does not mean that the multicast data is transferred through the 5G network only as a best effort service. The transparent multicast transport should utilize the policy and charging control framework which allows the interaction between PCF and AF. The multicast source, which takes the role of AF, can provide a filter information to identify the service data flows for policy control and/or differentiated charging. The multicast source can provide bandwidth requirements for QoS control of the flows [37].

6.1.3 Point-to-multipoint services

6.1.3.1 General

5G system may offer content delivery through a set of services, which may utilize the system's capability of point-to-multipoint data transport to a group of users or in a geographical area. We shall discuss the concepts used over xMB reference point to clarify what is meant by a service [13]. In the scope of the xMB specification [13], a service corresponds to a content provider's service offering for delivery over a network supporting MBMS to UEs. A deeper look to the specification reveals that a service is in fact a resource that holds a collection of sessions. Each session represents a delivery of content to UEs and can be of any of the following types: streaming, files, application, transport-mode. Consequently, a single service can be used to deliver files, streaming, application data or any data in transport-mode. The concept of services and sessions follows the legacy of MBMS architecture since Release 9 where a service in [13] corresponds to an MBMS User Service in [12] and a session in [13] resembles an MBMS User Service Session. It should be noted that both the service and the session are created by the BM-SC with default values and the content provider is responsible for updating them after the creation. A point-to-multipoint service refers to the service as defined in [13].

The BM-SC selects a delivery method for a session. The delivery methods are specified in [12] and includes the download delivery method, the streaming delivery method, the transparent delivery method (since Release 14) and the group communication delivery method (since Release 13). Although the specification does not define how the BM-SC shall select the delivery method, the choice of delivery method is rather intuitive. The download delivery method is suitable for the Files session type. A Streaming session will be most likely delivered using the streaming delivery method. Interestingly, DASH segments are delivered using the application session type for which the BM-SC could select the file download delivery method [13].

In MBMS, the BM-SC implements functionalities such as encapsulation and FEC for the download delivery method and the streaming delivery method. The BM-SC can also support the associated delivery procedures, e.g. a file repair procedure.

The xMB reference point could be adopted in 5G system including necessary functionalities and associated delivery procedures, which are discussed in the following sections, although underlying system functions for delivery of multicast and broadcast data (referred to as bearer services and bearer service architecture in [12]) may be different.

6.1.3.2 User data encapsulation and reliable content delivery

Delivering files or media segments using point-to-multipoint requires user data encapsulation techniques which differ from unicast delivery due to one-way communication between server and multiple clients. Indeed, efficient and reliable data delivery over a unidirectional and lossy channel implies the usage of dedicated data encapsulation protocols (e.g. FLUTE) and associated procedures. Unidirectional delivery protocols are designed to allow the use of AL-FEC. Furthermore, by associating the reliable data delivery with a repair procedure, a fully reliable delivery method can be achieved by allowing the UEs to request the missing data. The CDN could be used to spread the load meaning that the request/response repair messages are transmitted through the CDN instead of direct connection between the UEs and the original source. For the purpose of efficiency, a feedback mechanism can be used to efficiently adjust the appropriate protection level of AL-FEC. Besides the media and entertainment vertical, the file distribution service (e.g. file delivery method defined in [12]) is also required by practically all vertical use cases: V2X (e.g. software update, traffic message, etc.), IoT (e.g. massive software update) and for public safety (MCData file distribution

defined in [38], [39]). By using the xMB interface or its evolution in 5G, a content provider can provision the content to a file distribution service. A UE's application receives the content via an API such as File Delivery Application Service API from the MBMS Client (see 3GPP TS 26.347 [24]).

Regardless of whether one or multiple user plane network functions are required to provide the encapsulation and the forward error correction services, any user plane network function for point-to-multipoint services requires a control plane function(s) for configuration due to the CUPS principle. It seems logical to separate a control network function(s) related to unidirectional transport from other control network functions. Therefore, new control network function(s) for the control of the encapsulation and the forward error correction function(s) should be introduced in the 5G network.

6.1.3.3 Geographical broadcast

There are applications and services that require or benefit from the capability of delivering user data in a geographical area to users. Unlike in the cases when users (UEs) express explicitly an interest in receiving multicast user data (e.g. by joining IP multicast group) and the network can setup the resources based on this demand, the geographical broadcast requires that the network resources are allocated upon request from the content provider. The required functionality could be provided by a dedicated control network function (e.g. 5G-Xcast control network function described in section 7) responsible for geographical broadcast management, which then uses other network functions (e.g. SMF) responsible for configuration of network resources. It's noted that the geographical area could be extended to nation-wide for specific services (e.g. terrestrial broadcast TV services).

6.1.3.4 Audience size measurement and metric reporting

Starting from a given number (or higher) of UEs consuming the same content in a specific geographical area, broadcast delivery is more beneficial than point-to-point (i.e. unicast) delivery (e.g. in terms of spectrum efficiency). The ability to automatically switch between unicast and broadcast requires the measurement of the audience size.

The network side (in the domain of network operator or content provider) could explicitly ask the UEs to periodically send the reports including various metrics (e.g. QoE, network quality) or implicitly measures the audience size through different means (e.g. passive measurement). In the context of converged fixed and mobile network, this functionality is more sophisticated and is studied in the deliverables D4.2 "Converged Core Network" (to be published in September 2018) [25] and D4.3 "Session control and management" (to be published in December 2018) [25].

It's noted that the audience size measurements and metric reporting could also be exposed to the content provider. However, the exposed information is subject to privacy legislation.

6.1.3.5 Multicast Offloading

Multicast offloading refers to a capability of multicasting the popular or most requested content (both real time and non-real time). 3GPP MooD described in section 4.1 addressed the same issue.

For DASH and HLS, the content is made available at a variety of different bit rates (i.e. representations in DASH terminology) by providing alternative segments encoded at different bit rates covering aligned short intervals of playback time. UEs (more specifically, the middleware in the UEs) request segments from different representations according to the network conditions over unicast. Consequently, two UEs with different network conditions request different sets of segments. This precludes a multicast

offloading solution purely based on transparent multicast transport without any knowledge about the representation because the IP streams are different.

Multicast offloading for real time content is based on audience measurement. The network continuously analyses the consumption reports sent by the UEs to detect the popular live contents and trigger the multicast offloading to optimize the use of radio resource within the areas where the audience size is above a given threshold. For DASH content, the network selects and broadcasts a specific representation.

6.1.3.6 Other functionalities

Previous sections describe the important functionalities required for multicast and broadcast as a service. Other functionalities could also be used such as service announcement. Service announcement refers to the methods to announce the list of available services to the UE. For example, if the number of UEs needed to receive the broadcast content, the service announcement could be delivered through a broadcast channel instead of leveraging unicast delivery.

6.2 Existing Network Functions and Relevance to 5G-Xcast

6.2.1 UPF

We assume that PDU session connectivity service providing unicast data transport is always present at least for some use cases, e.g. multimedia and entertainment, and multicast is an integral part of the system used for optimization. Under this assumption, the unicast system architecture provides the foundation of 5G-Xcast architecture and the necessary architectural enhancements needed for the support of multicast should be aligned with the unicast architecture. However, the 5G-Xcast architecture's alignment with the unicast architecture should not require a unicast to be always present and thus prevent a network deployment (e.g. a network slice described in 5G-Xcast deliverable D4.2 "Converged Core Network") for cases when unicast may not be present, e.g. geographical broadcast of PW or to Receive-Only devices for terrestrial broadcast TV services. In Table 1, we review the UPF functionalities [21] and discuss their relevance in the respect to the introduction of point-to-multipoint support in the architecture.

Table 1: UPF functionalities for unicast and its comments for point-to-multipoint.

Functionality	Relevance to point-to-multipoint
Anchor point for Intra-/Inter-RAT mobility (when applicable).	It should remain the same.
External PDU Session point of interconnect to Data Network.	It should remain the same.
Packet routing & forwarding (e.g. support of Uplink classifier to route traffic flows to an instance of a data network, support of Branching point to support multi-homed PDU session).	Relevant when the point-to-multipoint source is a UE.
Packet inspection (e.g. Application detection based on service data flow template and the optional PFDs received from the SMF in addition).	Relevant in some solutions to PDU session procedures described in the deliverable D4.3 "Session Control and Management".
User Plane part of policy rule enforcement, e.g. Gating, Redirection, Traffic steering).	Using the same framework for point-to-point (i.e. unicast) and point-to-multipoint can simplify the architecture. The aim should be to reuse the policy rule enforcement functionalities of UPF as much as possible.

Lawful intercept (UP collection).	Relevant in the uplink when the point-to-multipoint source is a UE and in the downlink for point-to-multipoint when the network is aware of receiving UEs.
Traffic usage reporting.	Reports could be generated also for point-to-multipoint traffic.
QoS handling for user plane, e.g. UL/DL rate enforcement, Reflective QoS marking in DL.	Similar to the policy rule enforcement, the same framework for QoS handling at UPF simplifies the architecture.
Uplink Traffic verification (SDF to QoS Flow mapping).	Relevant when the source of multicast is a UE.
Transport level packet marking in the uplink and downlink.	It's considered as low priority in 5G-Xcast.
Downlink packet buffering and downlink data notification triggering.	UPF may transfer IP packets without buffering directly to downstream nodes in (R)AN). File delivery function requires buffering, other NFs could handle this functionality.
Sending and forwarding of one or more "end marker" to the source NG-RAN node.	This functionality is needed for switching N3/N9 tunnels during inter NG-RAN handover. This functionality may be relevant if the system supports mobility for point-to-multipoint.

6.2.2 SMF

The functionalities of SMF and their relevance to the support of multicast are reviewed in Table 2 [21].

Table 2: SMF functionalities for unicast and its comments for multicast.

Functionality	Relevance to point-to-multipoint
Session Management e.g. Session establishment, modify and release, including tunnel maintain between UPF and AN node.	To be enhanced to support point-to-multipoint.
UE IP address allocation & management (including optional Authorization).	A candidate for IP multicast address allocation, if IP multicast address are managed by 5GC.
DHCPv4 (server and client) and DHCPv6 (server and client) functions.	Not relevant.
Selection and control of UP function, including controlling the UPF to proxy ARP or IPv6 Neighbour Discovery, or to forward all ARP/IPv6 Neighbour Solicitation traffic to the SMF, for Ethernet PDU Sessions.	To be enhanced to support point-to-multipoint.
Configures traffic steering at UPF to route traffic to proper destination.	To be reused with minimal changes in comparison to unicast.
Termination of interfaces towards Policy control functions.	To be reused with minimal changes in comparison to unicast.
Lawful intercept (for SM events and interface to LI System).	No change.
Charging data collection and support of charging interfaces.	To be enhanced to support point-to-multipoint.

Control and coordination of charging data collection at UPF.	To be enhanced to support point-to-multipoint.
Termination of SM parts of NAS messages.	No change
Downlink Data Notification.	To enable activation of user plane connection. A service request procedure or a service announcement procedure could be triggered as described in the deliverable D4.3 “Session Control and Management”.
Initiator of AN specific SM information, sent via AMF over N2 to AN.	No change.
Determine SSC mode of a session.	The relevance will be discussed in the deliverable D4.3 “Session Control and Management”.

7 5G-Xcast Core Network Architecture

7.1 Introduction

This section describes three alternative architecture solutions that have been considered in 5G-Xcast. The network functions specific to multicast or broadcast are shown in the below figures. Common to all three alternatives is the functionalities in the UE which are not directly in scope of the current section. The three architecture alternatives are listed as follows:

1. Alternative 1 is described in section 7.2. This alternative provides an approach which is 5G architecture friendly.
2. Alternative 2 is described in section 7.3. This alternative provides a solution with minimal changes to the eMBMS architecture and specification described for LTE.
3. Alternative 3 is described in section 7.4. This alternative provides an approach where multicast and broadcast functionalities are integrated into the existing network functions.

7.2 Alternative 1

7.2.1 Overview

The 5G system architecture for alternative 1 is shown in Figure 17. In this alternative, the functionalities of BM-SC as they exist in LTE are split into a control plane part (XCF) and a user plane part (XUF). The XUF interfaces with the content provider via xMB-U interface and with the UPF via N6 reference point. The XUF provides multicast and broadcast specific functionalities for the user plane (see section 7.2.4).

The N6 reference point and the functionalities of UPF are enhanced to support and handle multicast traffic received from the XUF or the content source at the PDU session level, e.g. the UPF supports IGMP, MLD and PIM and the IP multicast routing is supported by the N6 interface for the IP PDU session type.

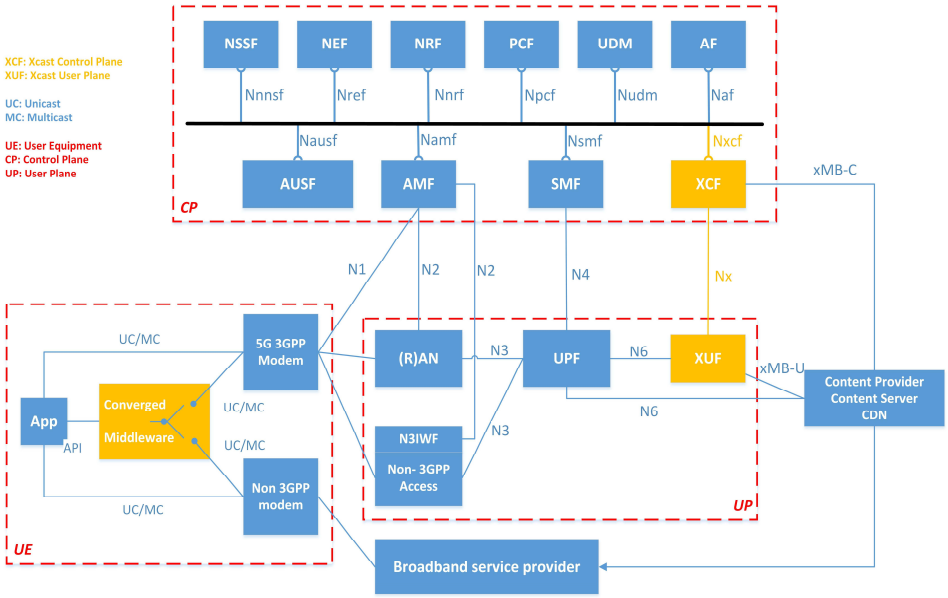


Figure 17. 5G System Architecture alternative 1.

7.2.2 UE functionalities

The UE functionalities described in this section apply to all three alternatives.

The UE architecture illustrates an abstract model of 5G-Xcast capable UE that consists of an application, a converged middleware, a 5G 3GPP modem and a non-3GPP modem. The modems offer connectivity services through access networks. The connectivity services provide exchange of PDUs between the Content Server and the modems and the XUF and the 5G 3GPP modem including multicast PDUs. The application may access the connectivity services directly, which typically involves with working network interfaces provided by an operating system. The role of the converged middleware is to hide the communication complexity to the application. The converged middleware could represent one of the following roles:

- Peer entity of XCF and XUF (i.e. similar to MBMS APIs in LTE [24]);
- Multicast GW of DVB multicast ABR as described in [40];
- HTTP client (e.g. available as a client library) with the support of multicast QUIC [41];
- ML-MW as described in the section 7.6.3.

It's noted that the converged middleware could be any combination of these roles described above. The functionalities of the converged middleware in the UE are specific to the services that the user consumes (e.g. the presentation of the BBC iPlayer is different from the presentation of a warning message) and to the deployment scenario (e.g. the functionalities of peer XCF and XUF entities are needed when XUF and XCF are deployed). The functionalities of the converged middleware in regard to the peer XCF and XUF entities are listed (non-exhaustive) as follows:

- Provisioning and configuration of user service parameters provided by XCF;
- Reception of user service announcement from XCF/XUF;
- Termination point of AL-FEC;
- Termination point of services offered by XCF/XUF (e.g. file delivery, multimedia content over multicast or broadcast, transparent transport of application data);
- Point-to-point file repair from XCF/XUF;
- Transmission of reception report and consumption report messages to the XCF/XUF;
- APIs to the applications.

When the application uses the PDU connectivity with multicast support described in the deliverable D4.3 "Session Control and Management" (to be published in December 2018) [25], an HTTP layer can hide the complexities of HTTP protocols such as redirections, version changes, underlying transport alternatives, etc. In the future, the HTTP layer may also support a concurrent multicast and unicast transport with HTTP over multicast QUIC [41].

The application may also use the services provided by the converged middleware via an API, which hides the complexities of unicast, multicast and broadcast transport to the application. The functionalities of the converged middleware as well as its peer entities in mobile core network (i.e. XCF, XUF) is expected to be standardized by 3GPP. In that case, the application will request a service from the converged middleware using the API [24] and the middleware will obtain the content from unicast, broadcast, or multicast using 5G 3GPP modem via 3GPP and non-3GPP radio access or any combination of these.

ML-MW is a part of the converged middleware which performs data combining of split data. Further information about the converged middleware is described in 5G-Xcast

deliverable D5.3 “Application Layer Intelligence” (to be published in December 2018) [25].

7.2.3 XCF functionalities

The XCF represents the peer endpoint to the content provider for the xMB-C reference point, i.e. the control plane part of xMB interface [12], [13]. The XCF functionalities related to xMB-C reference point includes the following:

- Authentication and authorization of XCF for a content provider;
- Authentication and authorization of a content provider for XCF;
- Creation, modification and termination of a service;
- Creation, modification and termination of a session;
- Status notification and query.

The XCF interacts with other network functions through service-based interfaces and uses the services offered by them to manage network resources for the xMB session. The following (non-exhaustive list) functionalities are supported in the XCF:

- Network resource management for xMB session using SMF services including
  - Allocation of UPF resources and maintenance core network tunnels between UPF(s) and (R)AN node(s);
  - Allocation of (R)AN resources by (R)AN upon SMF request(s) in the geographical area.
- AL-FEC configuration;
- Allocation of reference point for multicast data transport to the UE (i.e. a multicast IP address);
- Session and service announcement;
- Reception of consumption and reception report about a service;
- File repair management;
- Control multicast (or broadcast) transport availability based on the consumption reporting (i.e. functionality similar to 3GPP MooD in LTE);
- DRM (Digital Right Management) management;
- Multilink session setup and release upon request from UE;
- Estimation of QoS parameters for data transfer via each available link.

Most of these control functions are similar to the functions provided by the BM-SC in LTE. It’s noted that the current specification of xMB reference in 3GPP Release 14 may not fulfil the requirements for 5G multicast/broadcast capabilities and may need to be enhanced.

7.2.4 XUF functionalities

The XUF represents the peer endpoint for the content provider through the xMB-U reference point, i.e. the user plane part of xMB interface [12], [13]. The XUF functionalities related to xMB-U reference point include the following

- Delivery of content to XUF from the content provider;
- Retrieval of content by XUF from the content provider.

The XUF functionalities are the following (non-exhaustive list):

- Reliable delivery of data over unidirectional transport (e.g. FLUTE);
- AL-FEC to protect content against packet loss.

The XUF sends the multicast IP packets to the UPF over N6 reference point, which in turn sends the multicast IP packets via N3 tunnel.

7.2.5 UPF functionalities

The UPF includes the following functionalities in addition to the functionalities specified in 3GPP TS 23.501 [21]

- Multicast group membership discovery (e.g. IGMPv4 for IPv4 and MLD for IPv6);
- Multicast routing (e.g. PIM).

The 5G system uses a tunnel to deliver point-to-multipoint data from UPF to (R)AN. This tunnel can be a GTP-U tunnel using unicast IP same as the tunnel at N3 reference point for unicast data. When (R)AN is requested to establish a tunnel to UPF, (R)AN is also informed whether the tunnel is used for the transport of point-to-multipoint data. Alternatively, the tunnel could be using multicast IP (same as M1 interface in eMBMS), which can be beneficial if the same point-to-multipoint data are sent to large number of (R)AN nodes.

7.2.6 Analysis

The architecture follows the CUPS design principle in 5G. The architecture also separates the network functions used for offering multicast and broadcast services over xMB interface while the architecture also allows for the transparent multicast transport of data from the Content Server to the UPF using the multicast routing capabilities at the N6 reference point.

The architecture introduces new reference point Nx. The XCF controls the XUF over this reference point. The XUF can use this reference point to notify the XCF about its status and user plane events. An option in which the XUF connects to the SMF via the N4 reference point was considered. However, the N4 reference point between SMF and XUF would be only used for a transparent transport of messages between XCF and XUF. Hence, the introduction of a new reference point is more appropriate.

In this alternative, the UPF scalability potentially depends on the number of the multicast streams to be managed. However, the UPF scalability is an implementation issue which can be addressed in various ways. For example, one or more UPFs can handle a single multicast stream. A single UPF can handle both unicast and multicast or a single UPF can handle multiple multicast streams.

When a content is delivered to multiple UEs using point-to-multipoint service offered by the XCF and the XUF, the XUF retrieves or receives the data from the Content Server through xMB interface. The XUF then performs the encapsulation using FLUTE (or other encapsulation protocols) with the protection against packet loss using AL-FEC. The encapsulated data is sent over N6 reference point to the UPF as IP multicast packets. The UPF then forwards the encapsulated data through the N3 reference point to the (R)AN nodes. The (R)AN nodes transmit the encapsulated data to the UEs. The converged middleware decapsulates the encapsulated data with additional AL-FEC decoding if required and delivers the content to the application.

In cases where the XUF functionalities are not desirable (e.g. encapsulation, AL-FEC protection) or alternative solutions to the XUF functionalities are performed at the Content Server, the Content Server sends an IP multicast traffic to the UPF over the N6 reference point (i.e. network infrastructure between UPF and the Content Server must support multicast routing). This implies that XUF and XCF are not needed for such

deployments. The deployment scenarios in which the MNO may support the direct injection of IP multicast are the following (see the section 6.1.2). An example deployment scenario when the direct injection of IP multicast traffic is possible is the case of MEC deployment. Indeed, the MEC server sends the data to the UEs using LADN (Local Area Data Network) [21] which belong to the MNO. In this deployment, alternative solutions to the XUF functionalities may be implemented by a MEC application running on MEC host platform. Another example deployment scenario could utilise automatic multicast tunnelling (AMT) [42].

In this architecture, the N3 reference point shall support multicast in order to multicast the data packets to the appropriate (R)AN nodes. 3GPP in Release 15 does not support multicast at the N3 reference point.

This alternative has several points that need to be taken into account:

- The synchronization for multi-cell transmission cannot be performed in the XUF (like the SYNC protocol between the BM-SC and the eNodeB in LTE) since the N6 interface does not provide tunnelling functionality (e.g. GTP-U) for the SYNC protocol. However, the synchronization could be done at the RAN, which also avoids the need for the synchronization of the RAN with the core network entities.
- In this alternative, the UPF delivers the same content to multiple gNodeBs and in the case of a large number of gNodeBs (e.g. large geographic coverage) the support of multicast at the N3 reference point and possibly at the N9 reference point seems beneficial.

7.3 Alternative 2

7.3.1 Overview

The 5G system architecture for alternative 2 is shown in Figure 18. The alternative 2 differs from alternative 1 in the fact that in alternative 2 the XUF interfaces with the RAN directly (via an M1-NG reference point) whereas in alternative 1 the XUF interfaces with the RAN via the UPF. Hence, in alternative 2 the XUF needs to support generic UPF capabilities (needed for multicast or broadcast capabilities), whereas in alternative 1 the XUF would only require to support dedicated multicast or broadcast functionalities.

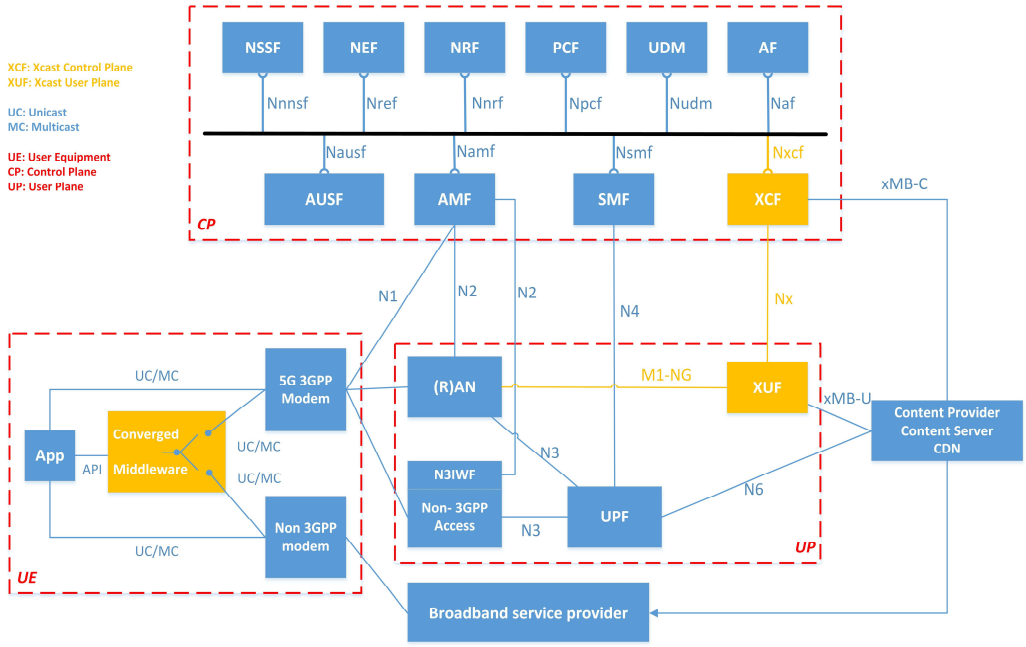


Figure 18. 5G System Architecture alternative 2.

An M1-NG interface is the same as M1 interface in LTE as specified in 3GPP TS 36.445 [18]. It uses GTP-U encoding to multicast the data packets to the RAN nodes. The XCF sends tunnel information IP multicast address to the RAN nodes or non-3GPP access nodes. From the protocol stack at IP and upper layers perspective, the N3 reference point in 5G use IP/UDP/GTP-U (see Figure 19). The difference is that M1-NG supports IP multicast while the N3 does not (up to 3GPP Release 15).

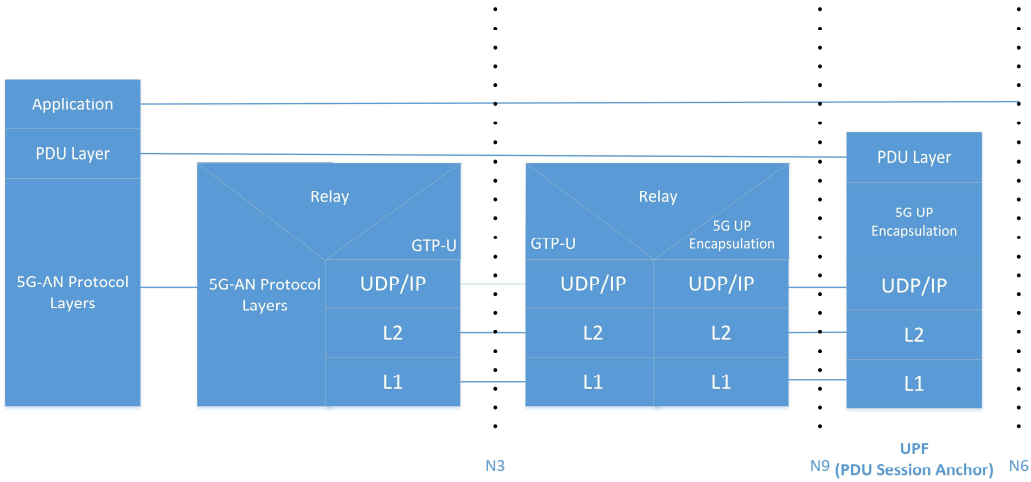


Figure 19. 5G user plane protocol stack.

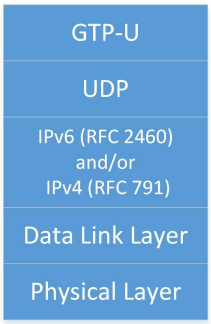


Figure 20. Transport network layer for MBMS data streams over M1 in LTE.

A unicast connection for file repair, reporting, etc. may require a point-to-point connection between the converged middleware and the network entities XCF and XUF. The reference point for this point-to-point connection is not shown in Figure 18.

7.3.2 Analysis

XUF implements a sub-set of UPF functionalities (e.g. packet routing and forwarding, traffic usage reporting, QoS handling, etc.) required for multicast or broadcast capabilities. The rationale is to offer minimized changes to the current eMBMS specification in LTE. Indeed, the architecture in this alternative is similar to the actual eMBMS architecture in LTE where the current BM-SC is split into control plane (e.g. XCF) and user plane (e.g. XUF). In this respect, the XUF in the alternative 1 is not an equivalent to the XUF in the alternative 2. The same applies to the XCF and its functionalities, which are different from the alternative 1. The XCF in the alternative 2 is responsible for a session management for point-to-multipoint services including the maintenance of tunnels and resources. The separation of control and user plane for BM-SC is already described in 3GPP TS 23.285 [43] (clause B.3 – option of localized user plane of MBMS Core Network function) which aims at the latency reduction for V2X services. Further description about localized MBMS is described in the Annex A.1.

This architecture alternative supports both transparent multicast transport (section 6.1.2) and point-to-multipoint as a service (i.e. broadcast functionalities in LTE) (section 6.1.3). Indeed, the solution where multicast functionalities are performed at the Content Server (e.g. through MEC platform) requires the multicast capabilities in UPF through N6 reference point as described in the alternative 1. In addition, multicast or broadcast capabilities as existed in LTE can be supported in 5G architecture through the XCF/XUF which reflect the BM-SC and MBMS-GW in LTE architecture. For instance, the XUF could provide SYNC functionality for a very large coverage area up to nation-wide as in LTE.

7.4 Alternative 3

7.4.1 Overview

The 5G system architecture for alternative 3 is shown in Figure 21. This alternative is based on alternative 1 and it supports both the transparent multicast transport via N6 reference point and the point-to-multipoint services offered via xMB interface. It should be noted that this alternative follows the principle that not every instance of UPF or SMF must support of all standardized functionalities as discussed in section 6.1.1. There are no changes to the transparent multicast transport in comparison to alternative 1. The changes in respect to the support of the point-to-multipoint services in comparison to

alternative 1 are the following. The XUF functionalities are part of UPF functionalities. The XCF functionalities could be split between the AF and the SMF.

The consequence of the UPF supporting the XUF functionalities is that the UPF may need to store the content if a functionality associated with a point-to-multipoint service requires, for example unicast file repair.

The AF is defined a generic network function for the interaction with 3GPP 5G core network. The AF implements xMB-C interface, which provides the same abstraction of point-to-multipoint services as in LTE. The AF can be responsible for multicast IP address allocation used by the services. In cases when there are multiple AFs implementing the XCF functionalities, the multicast IP address allocation needs to be coordinated for example by provisioning from the core network. The AF (XCF) controls the XUF functionalities associated with point-to-multipoint service and supported by the UPF (see section 7.2.4). The AF communicates directly or indirectly via the NEF with the SMF, which forwards control information to the UPF. The AF supports the service associated functionalities including consumption reporting, service announcement, etc. (see section 7.2.3) except of the network resource management.

The SMF is responsible for managing network resources for multicast or broadcast transport including the maintenance of core network tunnels between UPFs and (R)AN nodes in the same way as in alternative 1.

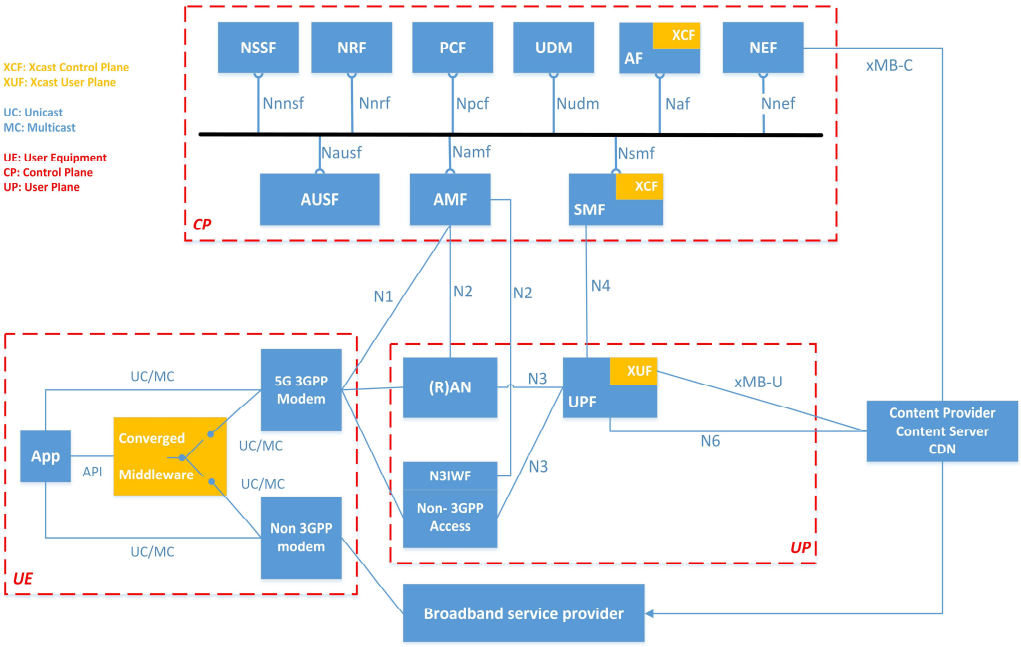


Figure 21. 5G System Architecture alternative 3.

In the alternative 3, the XUF is an integrated part of the UPF. The UPF interfaces with the data network via N6 reference point and the point-to-multipoint services are accessible via the xMB reference point.

7.4.2 Analysis

This architecture addresses the deployment challenges of alternative 1 when the N6 reference point between the UPF and the XUF cannot support IP multicast traffic (for e.g. the UPF and the XUF are not located in the same network), which is required for the point-to-multipoint services in the alternative 1. The issue is solved by enhancing the

UPF to provide the XUF functionalities. By doing so, it is possible to support the point-to-multipoint services with an infrastructure not supporting IP multicast routing. However, the alternative 3 does not exclude the option of supporting IP multicast routing at the N6 reference point, which allows for other deployments, e.g. local services.

The XCF takes the place of the AF in the architecture, which may raise a question whether the XCF as AF will be standardized by 3GPP. It is obvious that some level of standardization will be required. In general, the XCF operation may not be fully standardized. However, SMF services for controlling the XUF functionalities in the UPF and possibly SMF services for network resource management for multicast and broadcast will have to be standardized and make available directly or indirectly via NEF. One could also consider keeping the XCF as a “native” NF to the 5G core as in alternative 1 while moving the XUF functionalities to the UPF, which is seen as a valid option.

7.5 Analysis of mobile core network architecture alternatives

Sections 7.2, 7.3 and 7.4 describe three alternative solutions for the 5G-Xcast mobile core network architecture. All three architectures can deliver the same functionalities and support both transparent multicast transport (section 6.1.2) and point-to-multipoint services (section 6.1.3). The architectures are described with the reference to the 5G system architecture [21]. The architectures differ in the set of network functions (NFs), the functionalities each NF implements and the reference points between the NFs.

In the alternative 1, the functionalities needed to support the point-to-multipoint services are separated from the existing NFs in 5GC specified by 3GPP in Release 15. This solution leverages the transparent multicast transport introduced to the system by enhancing UPF, SMF, PCF, and possibly other NFs (e.g. UDSF). In addition, XCF and XUF are introduced to offer point-to-multipoint services. This alternative does not add new reference points to the 5GC except the Nx reference point between the XCF and the XUF. The N6 reference point needs to be enhanced to support IP multicast traffic. The N3 reference point may use the current point-to-point tunnelling for forwarding data to (R)AN nodes. Optionally, the N3 reference point could be enhanced to support point-to-multipoint tunnelling, which could be beneficial for delivery of data to a larger number of (R)AN nodes. The multi-cell transmission is being studied in the Work Package 3 of the 5G-Xcast project. In this alternative, the synchronization for multi-cell transmission cannot be performed in the XUF, however, it could be done at the (R)AN or the UPF. The advantage of the synchronization in the (R)AN is that the UPF does not need to be synchronized with the (R)AN. Moreover, the synchronization between nodes introduces a latency, which is assumed to be generally larger when the nodes are the (R)AN and the UPF due to differences between fronthaul and backhaul transport. On the other hand, the synchronization for multi-cell transmission in a large geographical area may become challenging.

The alternative 2 leverages the LTE eMBMS architecture for offering point-to-multipoint services. The functionalities needed to support these services are separated from the unicast functionalities and the transparent multicast transport. The functionalities are split into the control plane functionalities and the user plane functionalities supported by the XCF and the XUF, respectively. The functionalities of the XCF and the XUF in the alternative 2 are different from the ones in the other two alternatives and thus they are not their equivalents. In addition to the Nx reference point between the XCF and the XUF, this alternative introduces a new reference point M1-NG between the XUF and the (R)AN nodes. From the protocol stack perspective, the M1-NG reference point does not differ much from the M1 reference point between the MBMS-GW and the eNodeBs in the LTE

eMBMS architecture. This alternative allows for the synchronization for multi-cell transmission at the XUF. The alternative 2 aims at the minimization of the changes to the LTE eMBMS specification. The alternative 2 can also offer the transparent multicast transport through the same enhancements over the unicast architecture as in the cases of alternative 1 and alternative 3 The impacts of the segregation architecture for the unicast and the transparent multicast transport from the architecture for the point-to-multipoint on design, standardization and implementation of overall architecture have not been studied.

The alternative 3 does not introduce any new NFs instead the XUF functionalities are supported by UPF and the XCF takes a role of AF and possibly some XCF functionalities may need to be supported by SMF. The split of the XCF functionalities between XCF (as AF) and SMF has not been studied into the detail. Hence, it's not possible to conclude on the issue before the procedures over the core network architecture are studied in the deliverable D4.3 "Session Control and Management". If the studies will show that no XCF functionalities as per the alternative 1 will need to be implemented in the SMF then the alternative 3 can be understood as an implementation option of the alternative 1.

7.6 Analysis of Multilink in the mobile core network architecture

This section describes how the 5G-Xcast Multilink (ML) functionality is integrated with the relevant NFs in an exemplary one of the currently discussed three core network architecture alternatives. Three different types of relevant ML component are described as follows:

- ML-CP: additional functionality in the control plane of the mobile core network, which performs the estimation of QoS parameters for data transfer via each available link, multilink session setup and release;
- ML-UP: additional functionality in the user plane of the mobile core network, which performs data splitting, IP tunnel establishment;
- ML-MW: ML middleware functionality in the UE between the Application and the lower transport levels, which performs data combining, signalling (channel quality data transmitting), caching, providing ML session setup request (QoS parameters).

7.6.1 UPF ML

The following control functions are included in the UPF:

- Data split;
- Establishment of IP tunnels.

7.6.2 SMF ML

The following control functions are included in the SMF:

- Multilink session setup and release.

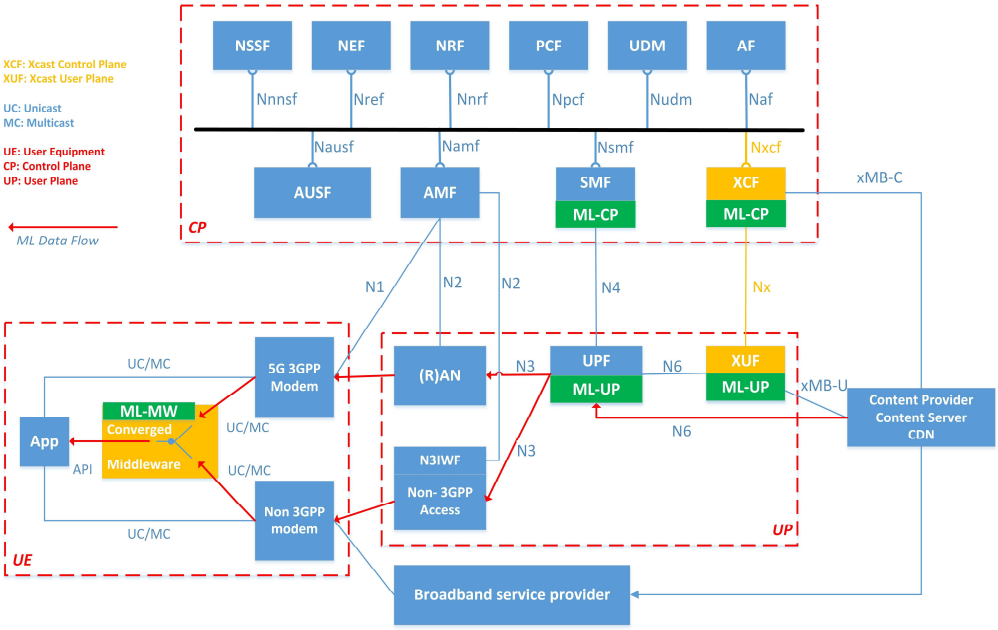
7.6.3 MW-ML

The following control functions are included in the converged middleware:

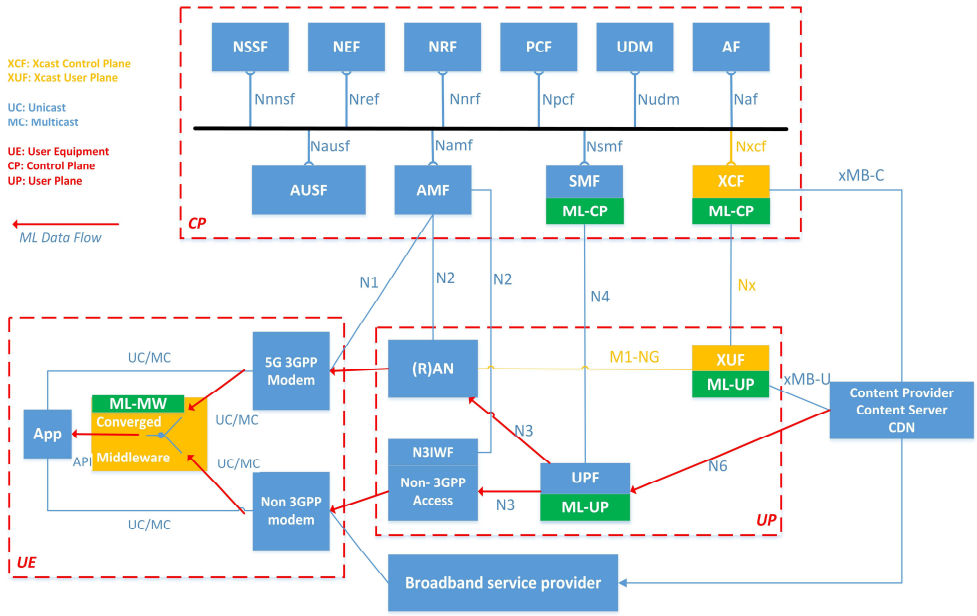
- Data combining;
- ML session Setup Request (QoS parameters);
- Signalling (channel quality data transmitting);
- Caching.

7.6.4 ML operations and analysis

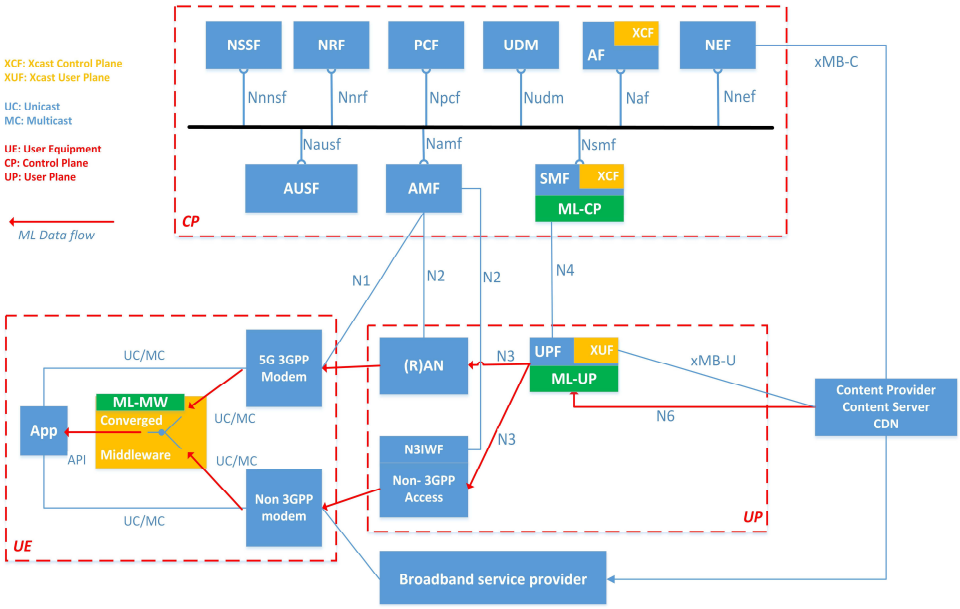
Figure 22 shows an example of data flow and the process of data transfer using ML. There is one common data stream from Content Server to UPF via the N6 reference point. The UPF then splits the data flow between 3GPP access and non-3GPP access. After receiving at the converged middleware, the split data is combined by ML-MW into one common stream. It's noted that data split will be provided by UPF and data combining by ML-MW for each architecture alternative while the difference will be at control plane functions. The ML integration in the core network architecture was done in a way that makes it optional and without loading the core architecture itself.



(a)



(b)



(c)

Figure 22. Multilink data flows in 5G-Xcast architecture (a – Alternative 1; b – Alternative 2; c – Alternative 3).

7.6.4.1 Mapping between Multilink and on-going work in 3GPP

This section describes how the 5G-Xcast core network with Multilink is mapped with the on-going work in 3GPP on "Study on Access Traffic Steering, Switch and Splitting support in the 5G system architecture" (3GPP TR 23.793 [44]).

The status of Wi-Fi Access integration into the 5G core network described in TS 23.501 is as follows:

- Connectivity of the UE only via non-3GPP access networks (e.g. WLAN access) is allowed.
- Non-3GPP access is standalone untrusted non-3GGP access.
- Non-3GPP access networks are connected to 5G core network via a Non-3GPP InterWorking Function (N3IWF):
    a) A UE shall establish an IPSec tunnel with the N3IWF, the UE shall be authenticated by and attached to the 5G Core Network during the IPSec tunnel establishment procedure.
    b) N1 NAS signalling over standalone non-3GPP accesses shall be protected with the same security mechanism applied for N1 over 3GPP access.
- A UE access the 5G Core Network supports NAS signalling using N1 reference point
- The N3IWF interfaces to 5G core network control-plane functions and user-plane functions via N2 interface and N3 interface, respectively.
- Only Untrusted Non 3GPP Access is currently in scope, Trusted being discussed for Release 16

All solutions considered in 3GPP TR 23.793 are based on and will be aligned with the 5GC Phase-1 normative work including work on policy management, as mentioned in TS 23.501 [21], TS 23.502 [22] and TS 23.503 [37].

In particular, 3GPP TR 23.793 considers the solutions that specify the following:

- How the 5GC and the 5G UE can support multi-access traffic steering between 3GPP and non-3GPP accesses;
- How the 5GC and the 5G UE can support multi-access traffic switching between 3GPP and non-3GPP accesses;
- How the 5GC and the 5G UE can support multi-access traffic splitting.

The scope of the document 3GPP TR 23.7893 excludes the following aspects:

- Changes to the charging framework are not considered. However, it may be considered what information needs to be provided to the charging framework in order to charge traffic that is switched and/or split between 3GPP and non-3GPP accesses;
- ATSSS procedures that may be applied in the NG-RAN are not considered. The study is restricted only to ATSSS procedures applied in the 5G core network.
- 5GS enhancements to support trusted non-3GPP access networks are not considered;
- 5GS enhancements to support wireline access networks are not considered.

The study in 3GPP TR 23.793 is organized into two phases. Currently, the study considers ATSSS solutions that enable traffic selection, switching and splitting between NG-RAN and untrusted non-3GPP access networks. And only after the 5GC architecture is enhanced to support trusted non-3GPP access networks, the study will also consider ATSSS solutions that enable traffic selection, switching and splitting between NG-RAN and trusted non-3GPP access networks.

Comparing to the solutions provided in the 3GPP technical report mentioned above, Multilink gives more freedom and possibilities to integrate 3GPP and non-3GPP (trusted and non-trusted) accesses into the 5G architecture (e.g. Wi-Fi, wireline, satellite etc.). As described in section 4.3.2, there are several strategies of using multiple links within 5G-Xcast project, depending of external circumstances: replication of the content, switching between links, load balancing, complementing via additional links and bonding and aggregation. For each exact use case it will bring more benefits than not yet fully defined 3GPP solutions.

8 Conclusions

In this deliverable, a thorough analysis of LTE eMBMS architecture has been performed, which revealed several shortcomings of the eMBMS architecture. The LTE eMBMS architecture lacks particularly of the dynamicity of MBSFN configuration and provisioning feedback regarding MBMS session management also including feedback from RAN to the core network. Mechanisms for UE triggering to receive MBMS have been found inefficient for services such as public warning. The service continuity between MBSFN areas has also been identified as an issue. In addition, the deliverable explains why the cell broadcast existing functionalities in 3GPP cannot support the multimedia public warning alert use case as described in the deliverable D2.1 [2].

This document also presents a new concept of converged autonomous switch between unicast/multicast/broadcast not only for mobile networks but also between access networks including fixed networks. Other emerging technologies relevant to the 5G-Xcast project are multi-connectivity and multilink technologies, and multi-access edge computing.

In heterogeneous networks, multi-connectivity helps to provide an optimal user experience (for e.g. high bandwidth, network coverage, reliable mobility). Multilink is one type of multi connectivity technology which refers to IP link aggregation where the IP links represent connectivity via the same or different access technologies offered by one or more operators. Multilink uses methods such as replication, switching, load balancing and aggregation to improve aspects of user's connectivity including overall bandwidth, reliability and availability, and mobility. An analysis on how multilink may be added to the 5G-Xcast mobile core network for the purpose of serving the broadcast/multicast transition with unicast has been performed. For each described alternative multilink functionality is integrated into several NFs.

Multi-access edge computing (MEC) is a key technology of 5G networks. In the MEC paradigm, IT and telecommunications converge into a new system where cloud computing is a part of the telecommunication network that offers computing, storage and network resources at various network locations typically including several edge network locations. MEC can offer necessary capacity and latency to support virtual and augmented reality services as documented in the deliverable D2.1 for the Media & Entertainment use case for virtual and augmented reality broadcast (M&E #2) [2]. MEC can also play an important role in delivering multimedia content in general, including the hybrid broadcast service main use case (M&E #1), as content hosting nearer to users can reduce CAPEX investments to transport networks of MNOs.

We defined design principles to guide the 5G-Xcast system design. The design principles follow the 3GPP principles of the 5G system architecture. The design principles, the analysis of eMBMS and the introduction of new technologies in the telecommunication field as well as in the IT domain were the inputs to a discussion on the modularization and function separation for multicast and broadcast. We considered various approaches to the problem and concluded that the 3GPP 5G system should allow for a transparent multicast transport and it should also offer point-to-multipoint services. In the case of transparent multicast transport, the integration of the 5G system with third-party systems is done at PDU session level in the user plane, possibly with an interaction between the control plane of 5G system and the third-party systems via the service-based interface of 5GC directly or indirectly, for example, to allow the third party to provision QoS

information. The examples of a third-party system are DVB mABR [40] and any system utilizing HTTP over multicast QUIC as described in the deliverable D5.2 “Key Technologies for Content Distribution Network”. The point-to-multipoint services are proposed to be offered via the xMB reference point as the 5G-Xcast did not identify any issues with the xMB reference point based on the feedback from the project partners with the view of a broadcaster. This topic will be studied in the second half of the project.

Finally, two primary alternative solutions for the 5G-Xcast mobile core network architecture are documented in this deliverable. Both architectures can deliver the same functionalities and support both transparent multicast transport (section 6.1.2) and point-to-multipoint services (section 6.1.3). The architectures are described with reference to the 5G system architecture [21] and differ in the set of network functions (NFs), the functionalities each NF implements and the reference points between the involved NFs.

In alternative 1, the functionalities needed to support the point-to-multipoint services are separated from the existing NFs in the 5GC specified by 3GPP in Release 15. Alternative 1 leverages the transparent multicast transport introduced to the system by enhancing UPF, SMF, PCF, and possibly other NFs (e.g. UDSF). In addition, two new 5G-Xcast related network functions (i.e. XCF and XUF) are introduced to offer point-to-multipoint services. This alternative does not add new reference points to the 5GC except the Nx reference point between the XCF and the XUF. The N6 reference point needs to be enhanced to support IP multicast traffic. In this alternative, the synchronization for multi-cell transmission cannot be performed in the XUF, however, it could be done at the (R)AN or the UPF.

Alternative 2 leverages the LTE eMBMS architecture for offering point-to-multipoint services. The functionalities needed to support these services are separated from the unicast functionalities and the transparent multicast transport. In addition to the Nx reference point between the new 5G-Xcast related network functions XCF and XUF, this alternative introduces a new reference point M1-NG (similar to the M1 reference point in the LTE eMBMS architecture) between the XUF and the (R)AN nodes. This alternative allows for the synchronization for multi-cell transmission at the XUF. Alternative 2 aims at minimizing the changes to the functionalities developed in LTE eMBMS. Alternative 2 can also offer the transparent multicast transport through the same enhancements over the unicast architecture as in the case of alternative 1.

As a potential secondary solution, this deliverable also considers another possible alternative that does not introduce any new NFs. However, in this secondary solution, the XUF functionalities are supported by the UPF and the XCF takes the role of AF and possibly some XCF functionalities may need to be supported by SMF. The split of the XCF functionalities between XCF (as an AF) and SMF has not been studied in detail. Hence, it’s not possible to conclude on the issue before the procedures over the core network architecture are studied in deliverable D4.3 “Session Control and Management”. If the studies show that no XCF functionalities as per alternative 1 will need to be implemented in the SMF then this possible alternative could be seen as an implementation option of alternative 1 at the time of writing this document.

The proposed architecture alternatives in this Work Package 4 (WP) will be a basis for other tasks in the same WP as well as other WPs in the 5G-Xcast project. More specifically, the work in deliverable D4.2 will take into account the proposed alternatives to build the converged network including fixed broadband, mobile and possibly broadcast networks while deliverable D4.3 will study the detailed work flows from the session

control and management perspective. Finally, this document intends to become a reference point for future standardization work in 3GPP Technical Specification Group SA2 to integrate multicast and broadcast in the 5G core network.

8.1 Open topics

This document describes the work performed during the first half of the project. The following open topics will be addressed in the second half of the project:

- Evaluation of ML depends on the on-going ATSSS work in 3GPP. At the time of writing this deliverable, the conclusion is that ML is better than ATSSS.
- AL-FEC for low latency will be studied jointly with Work Package 3 and will be described in the deliverable D3.4 “RAT Protocols and Radio Resource Management”.
- Synchronization for multi-cell transmission is being studied in Work Package 3, especially in deliverable D3.3 “RAN Logical Architecture” (to be published in March 2019) [25]. The outcome of WP3 will be considered and the architectures updated, if needed.
- The UE functionalities are being studied in deliverable D5.3 “Application Layer Intelligence”.
- The applicability of the proposed core architectures for terrestrial broadcast will be studied in the second half of the project.
- The mobile core network architecture alternatives described in the present document will be extended to the converged network including both fixed broadband and mobile networks. This work is being studied in the deliverable D4.2 “Converged Core Network”.
- The work flows and procedures based on the architecture alternatives are being studied in the deliverable D4.3 “Session Control and Management”.
- Security issues potentially arising with the chosen architecture/solution(s) will need to be properly addressed.

References

[1] 3GPP TR 22.261 v16.3.0, “Service requirements for next generation new services and markets,” March 2018.

[2] D. Ratkaj and A. Murphy, Eds., “Definition of Use Cases, Requirements and KPIs,” Deliverable D2.1, 5G-PPP 5G-Xcast project, Oct. 2017.

[3] 3GPP RP-180499: “Interim report from email discussion on 5G Broadcast evolution”.

[4] 3GPP news, “Enhanced Television Services over 3GPP eMBMS,” Oct. 2017. http://www.3gpp.org/news-events/3gpp-news/1905-embms_r14

[5] 3GPP TR 38.913 v14.3.0, “Study on Scenarios and Requirements for Next Generation Access Technologies,” June 2017.

[6] D. Gomez-Barquero, D. Navratil, S. Appleby and M. Stagg, “Point-to-Multipoint Communication Enablers for the Fifth-Generation of Wireless Systems”, IEEE Communications Standards Magazine, vol. 2, no. 1, pp. 53-59, March 2018.

[7] N. Nouvel, Ed., “Content Delivery Vision,” Deliverable D5.1, 5G-PPP 5G-Xcast project, Nov. 2017.

[8] D. Vargas and D. Mi, Eds., “LTE-Advanced Pro Broadcast Radio Access Network Benchmark,” Deliverable D3.1, 5G-PPP 5G-Xcast project, Nov. 2017.

[9] 3GPP TS 36.300: "Evolved Universal Terrestrial Radio Access (E-UTRA) and Evolved Universal Terrestrial Radio Access Network (E-UTRAN); Overall description; Stage 2".

[10] 3GPP TS 23.246: "Multimedia Broadcast/Multicast Service (MBMS); Architecture and functional description".

[11] 3GPP TS 25.446: “MBMS synchronization protocol (SYNC)”.

[12] 3GPP TS 26.346: "Multimedia Broadcast/Multicast Service (MBMS); Protocols and codecs".

[13] 3GPP TS 29.116: “Representational state transfer over xMB reference point between content provider and BM-SC”.

[14] 3GPP TS 23.468: “Group Communication System Enablers for LTE (GCSE_LTE); Stage 2”.

[15] 3GPP TS 29.468: "Group Communication System Enablers for LTE (GCSE_LTE); MB2 Reference Point; stage 3".

[16] 3GPP TS 29.061: "Interworking between the Public Land Mobile Network (PLMN) supporting packet based services and Packet Data Networks (PDN)".

[17] 3GPP TS 29.274: “3GPP Evolved Packet System (EPS); Evolved General Packet Radio Service (GPRS) Tunnelling Protocol for Control plane (GTPv2-C); Stage 3”.

[18] 3GPP TS 36.445: "Evolved Universal Terrestrial Radio Access Network (E-UTRAN); M1 data transport".

[19] 3GPP TS 36.443: “Evolved Universal Terrestrial Radio Access Network (E-UTRAN); M2 Application Protocol (M2AP)”.

[20] 3GPP TS 36.444: “Evolved Universal Terrestrial Radio Access Network (E-UTRAN); M3 Application Protocol (M3AP)”.

[21] 3GPP TS 23.501: “System Architecture for the 5G System”.

[22] 3GPP TS 23.502: “Procedures for the 5G System”.

[23] 3GPP TR 26.849: “Multimedia Broadcast/Multicast Service (MBMS) improvements; MBMS operation on demand”.

[24] 3GPP TS 26.347: “Multimedia Broadcast/Multicast Service (MBMS); Application Programming Interface and URL”.

[25] 5G-Xcast documents, http://5g-xcast.eu/documents/

[26] 3GPP TS 23.041: “Technical realization of Cell Broadcast Service (CBS)”.

[27] ATIS-0700008.v002: “Cell Broadcast Entity (CBE) to Cell Broadcast Center (CBC) Interface Specification, Revision 2”.
[28] 3GPP TS 29.168: “Cell Broadcast Centre interfaces with the Evolved Packet Core; Stage 3”.
[29] 3GPP TS 36.413: “Evolved Universal Terrestrial Radio Access Network (E-UTRAN); S1 Application Protocol (S1AP)”.
[30] ETSI GS MEC 003: “Mobile Edge Computing (MEC); Framework and Reference Architecture”.
[31] ETSI GS MEC 011, V1.1.1, “Mobile Edge Computing (MEC); Mobile Edge Platform Application Enablement,” July 2017.
[32] ETSI GS MEC 012, V1.1.1, “Mobile Edge Computing (MEC); Radio Network Information API,” July 2017.
[33] ETSI GS MEC 013, V1.1.1, “Mobile Edge Computing (MEC); Location API,” July 2017.
[34] ETSI GS MEC 015, V1.1.1, “Mobile Edge Computing (MEC); Bandwidth Management API,” October 2017.
[35] 3GPP TS 23.379: “Functional architecture and information flows to support Mission Critical Push To Talk (MCPTT); Stage 2”.
[36] 3GPP TS 23.280: “Common functional architecture to support mission critical services; Stage 2”.
[37] 3GPP TS 23.503: “Policy and Charging Control Framework for the 5G System; Stage 2”.
[38] 3GPP TS 22.282: “Mission Critical Data over LTE”.
[39] 3GPP TS 23.282: “Functional architecture and information flows to support Mission Critical Data (MCData); Stage 2”.
[40] https://www.dvb.org/news/dvb-releases-reference-architecture-for-ip-multicast
[41] https://datatracker.ietf.org/doc/draft-pardue-quic-http-mcast/
[42] IETF RFC7450, “Automatic Multicast Tunneling”, February 2015.
[43] 3GPP TS 23.285: “Architecture enhancements for V2X services”.
[44] 3GPP TR 23.793: “Study on Access Traffic Steering, Switch and Splitting support in the 5G system architecture”.
[45] 3GPP TS 23.003: “Numbering, addressing and identification”.
[46] 3GPP S6-171504 http://www.3gpp.org/ftp/TSG_SA/WG6_MissionCritical/TSGS6_020_Reno/Docs/S6-171504.zip
[47] 3GPP TR 23.780: “Study on Multimedia Broadcast and Multicast Service (MBMS) usage for mission critical communication services”.
[48] Hong Wang, H. Vandervelde and S. Kim, "LTE MBMS SYNC protocol for support synchronization of content," *2009 IEEE International Conference on Communications Technology and Applications*, Beijing, 2009, pp. 392-395.
[49] 3GPP TR 23.780: "Study on Multimedia Broadcast and Multicast Service (MBMS) usage for mission critical communication services".
[50] https://en.wikipedia.org/wiki/Xcast

A Annex

A.1 Analysis of 3GPP eMBMS Release 14 limitations

This section discusses about the limitations of 3GPP eMBMS Release 14. Hence, the terminologies in this section are used based on the view from 3GPP. In addition to the limitations described in the following sections, LTE lacks of convergence in the architecture or implementation in industry. The lack of convergence motivates the work developed in deliverable D4.2 “Converged Core Network” (to be published in September 2018) [25].

A.1.1 Lack of dynamic configuration for MBSFN

A.1.1.1 On MBMS Service Area

MBMS Service Area and MBSFN Area are statically preconfigured. The cells belonging to MBSFN Areas must be configured before it is possible to launch an MBMS bearer. Inside the MBMS bearer context parameters, a field called “List of Cell ID(s)” alongside “MBMS Service Area”, defines the region where the MBMS service may be delivered (see 3GPP TS 23.246 [10]). While there are mechanisms that allow the BM-SC to modify the list of MBMS Service Areas for an on-going MBMS service (see 3GPP TS 29.061 section 20.3.2 [16]) for MBMS session, the MBMS Service Areas themselves are statically configured by the MNO and are not adaptive to user demand. The cell list improves the dynamicity of MBMS Service Area definition, but it still requires the OAM configuration of MBMS Service Areas.

3GPP MBMS operation on Demand (MooD) [12], [23] is impacted by the static configuration of MBMS Service Area. 3GPP MooD allows seamlessly switching between unicast data bearers and MBMS data bearer based on user traffic but is restricted to a MBMS Service Area. Depending on the static configuration of MBMS Service Area, the unicast-to-multicast traffic offloading may initially occur in a MBMS Service Area covering a wide range of cells where the demand may only come from one cell. Either a 1:1 mapping of SAI to a cell or the use of the cell list introduced as part of SC-PTM is needed to start an MBMS session in the interested cells. Currently MBMS SAI assignment to cells is achieved through OAM configuration.

In addition, within the same MBMS session, there is no available procedure to quickly switch between SC-PTM and MBSFN delivery mode based on user traffic without relaunching the MBMS bearer. Relaunching the MBMS bearer could cause interruption of the service and force the UE to reconnect to the new MBMS session.

Since the air interface channels used for SC-PTM and MBSFN are different, extending SC-PTM over one cell to form a MBSFN Area is impossible, without relaunching the MBMS Bearer to change the delivery mode to MBSFN. Even if this procedure is manually performed by OAM, there are no specified protocol to inform the UE of this change.

It’s also noted that MBSFN and SC-PTM delivery modes use different physical channels, hence, these delivery modes are chipset dependent. A UE that supports one mode may not support the other mode, hence, switching back and forth between these modes may not be possible at all for a UE.

3GPP MooD is not able to switch between SC-PTM and MBSFN, even if one mode is more optimal for media delivery over the other in a certain scenario. An ideal system should switch between unicast, to Single Cell point-to-multipoint, up to a dynamically created MBSFN Area, based on user demand, and vice versa.

A.1.1.2 On Distributed MCE deployment

In a distributed MCE deployment where MCE is collocated with eNodeB, forming a MBSFN area spanning multiple MCEs/eNodeBs needs direct intervention of OAM to ensure that the operation of each involved MCE is coordinated. An OAM interface with every MCE/eNodeB is needed to ensure the MBSFN area configuration is the same across all cells and MCEs/eNodeBs. This interface is implementation specific to a network vendor which makes multi-vendor deployment of larger MBSFN areas more challenging. Using a standard interface, which interconnects RAN nodes such as the X2 interface interconnecting eNodeBs, to enable this coordination process can be explored.

More specifically, when the BM-SC establishes an MBMS session with the desired QoS, every MCE/eNodeB forming the MBSFN Area, which receives the MBMS session start request message, must configure the exact same RAN parameters of MBSFN area (MCCH, MTCH) to enable the SFN. The MNO must consider the implications of a coordinated configuration of air interface parameters for every eNodeB in the MBSFN Area, especially if there is a diverse set of eNodeB hardware and software used on the eNodeBs forming the SFN with its own MCE implementation.

A.1.1.3 On Service Area Identifier

SAI (Service Area Identifier) is an identifier that references a single or a group of cells. Its value ranges from 0 to 65,535 [45]. A 1:1 SAI to cell mapping is not practical because the number of cells in the network may be significantly larger than the maximum number of 65,535. While nation-wide MBMS can be enabled using the special value “0”, it is still not practical to achieve this in case of MNO sharing infrastructure or when specific service (e.g. PWS) requires fast setup of region-wide MBMS across different MNOs and hence MBMS configuration between different MNOs would be required in this case. Another issue with 1:1 SAI to cell mapping is the possible loss of advantage of MBSFN delivery mode due to MBMS Service Area containing a single cell (see Figure 3 in section 2.1.2.1). In that case, SC-PTM delivery mode is seen as more appropriate.

However, there is an interest to map each SAI value to an individual cell, in order to have access to eMBMS transmission with cell granularity (e.g. Public Warning Services). LTE addressed the issue by using cell IDs (ECGIs) along with MBMS SAI to provide cell granularity of the MBMS service area. The issue is that BM-SC still must know the correct MBMS SAI because MBMS SAIs are used to route the control messages towards downstream nodes all the way up to MCE, which then uses the cell IDs to configure the broadcast areas. This two-layer approach increases system’s complexity and OAM effort.

Therefore, it may be beneficial to integrate SC-PTM and MBSFN in a single PTM (Point to Multipoint) framework.

A.1.2 Lack of feedback on eMBMS session management

Session management is used to Start/Stop an eMBMS session or to modify (update) parameters of an on-going eMBMS session. The current eMBMS session management faces the following challenges:

- When an eMBMS session is requested by a BM-SC, the BM-SC has no knowledge if the session has actually started in each eNodeB;
- During an ongoing eMBMS session, the BM-SC has no knowledge if an eMBMS session is pre-empted (terminated) in an eNodeB

It is possible for eMBMS sessions to fail in the eNodeB due to various reasons such as insufficient capacity, configuration error, software or hardware fault, connectivity

problem, or a pre-emption vulnerable bearer due to a bearer with a higher priority. The following sub-sections identify and analyse the limitation on different interfaces.

A.1.2.1 On SGmb interface

To start an eMBMS session, the BM-SC performs the MBMS Session Start procedure and sends a Re-Auth-Request (RAR) command with MBMS-StartStop-Indication to the MBMS-GW. In response to the request from BM-SC, the MBMS-GW sends the Re-Auth-Answer (RAA) command with MBMS-StartStop-Indication. Figure 23 shows the MBMS Start procedure as described in section 8.3.2 of 3GPP TS 23.246 [10].

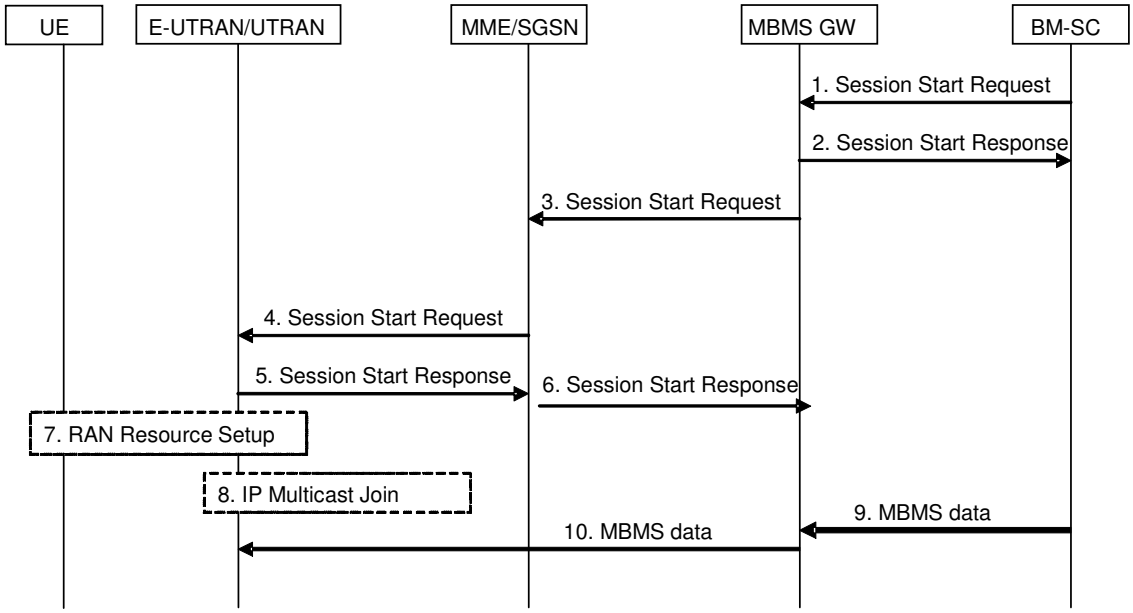


Figure 23: MBMS Start Procedure from 3GPP TS 23.246.

As shown in Figure 23, the MBMS-GW may not need to wait for a response from the network (MME) but it sends the Session Start Response to the BM-SC based on the success of its own local operation related to the MBMS session start procedure, i.e. context creation and SGi-mb user plane allocation at MBMS-GW (Message 2 in Figure 23), and therefore the MBMS-GW has the sole responsibility to start the session [46]. However, the 3GPP specification [10] in Release 15 also allows for implementations when MBMS GW could wait for the response from the network (see Figure 8b.2 of [10]).

As per 3GPP TS 23.246 [10], the supported procedures are:

- Session Start;
- Session Stop;
- Session Update (only invoked by BM-SC to change parameters, not invoked by MBMS-GW);
- Heartbeat – keep alive messages

There is no procedure to update the BM-SC as result of changes in eMBMS sessions running in the network, for example MBMS session pre-emption. In addition, there is also no message defined in SGmb to support this function.

A.1.2.2 On Sm interface

Table 3 gives the impression that the MME is capable of providing feedback to the MBMS-GW, as message 6 “Session Start Response” is drawn slightly after message 5 “Session Start Response” from MCE / eNodeB (Figure 23).

The message 6 “Session Start Response” from MME to MBMS-GW supports the following IE (Information Element) as shown in Table 3, taken from 3GPP TS 29.274 section 7.13.2 [17]:

Table 3: Sm Session start response IE.

Information elements	P	Condition / Comment	IE Type	Ins.
Cause	M		Cause	0
Sender F-TEID for Control Plane	M		F-TEID	0
MBMS Distribution Acknowledge	C	This IE shall be included on the Sn interface.	MBMS Distribution Acknowledge	0
Sn-U SGSN F-TEID	C	This IE shall be included on the Sn interface if some RNCs have not accepted IP multicast distribution.	F-TEID	1
Recovery	C	This IE shall be included if contacting the peer for the first time.	Recovery	0
Private Extension	O		Private Extension	VS

There is only a single Cause field, even though the MME is likely to address multiple eNodeB/MCE for a single eMBMS session. Hence, there is currently no possibility to provide detailed information on which eNodeB a session started or failed using this interface.

An eMBMS session start is deemed to be successful if the service started in one eNodeB but failed to start in all others.

There is currently no procedure or message for the MME to update the MBMS-GW with changed information on running eMBMS sessions, for example MBMS session pre-emption.

A.1.2.3 On M3 interface

The M3 interface provides session management between the MME and MCE. The MCE can be either centralized or distributed as described in section 2.1. Table 4 shows the list of functions and procedures taken from 3GPP TS 36.444 [20].

Table 4: Mapping between M3AP functions and M3AP EPs.

Function	Elementary Procedure(s)
Session Management	a) MBMS Session Start b) MBMS Session Stop c) MBMS Session Update
Error Indication Functionality	Error Indication
Reset Functionality	Reset
M3 Setup	M3 Setup
Configuration Update	MCE Configuration Update

In case of distributed MCE architecture where MCE and eNodeB has 1:1 relationship, the MME could determine in which eNodeB sessions are started.

In case of centralized MCE architecture, the response of the MBMS Session Start can only have one Cause code, as can be seen from Table 5.

Table 5: MBMS Session Start Failure IE.

IE/Group Name	Presence	Range	IE type and reference	Semantics description	Criticality	Assigned Criticality
Message Type	M		9.2.1.1		YES	reject
MME MBMS M3AP ID	M		9.2.3.2		YES	ignore
Cause	M		9.2.1.2		YES	ignore
Criticality Diagnostics	O		9.2.1.7		YES	ignore

With just one Cause code, different responses from multiple eNodeBs cannot be signalled from MCE to MME. This also limits the MCE to provide information on what method (SC-PTM or MBSFN) was chosen for the transmission to other nodes which may be relevant in scenarios when UE does not support both transmission modes. No message has been defined for the MCE to report any issues with an eMBMS session after it has started.

A.1.2.4 On MB2-C interface

The MB2-C interface provides session management between the GCS AS and BM-SC for MCPTT service. The GCS AS requests an MBMS bearer through the GCS-Action-Request (GAR) and the BM-SC responds with GCS-Action-Answer. The GCS-Action-Answer contains the response for the MBMS bearer in the AVP MBMS-Bearer-Response which contains of the following AVP:

```
MBMS-Bearer-Response::=          < AVP Header: 3505 >
                                 [ TMGI]
                                 [ MBMS-Flow-Identifier ]
                                 [ MBMS-Session-Duration ]
                                 [ MBMS-Bearer-Result ]
                                 [ BMSC-Address ]
                                 [ BMSC-Port ]
                                 [ MB2U-Security ]
                                *[ Radio-Frequency ]
                                *[ AVP ]
```

For the MBMS-Bearer-Result AVP, any value other than zero is a failure, a partial failure has not yet been defined (but could be done in theory). An additional AVP could be added to also define in which SAI/Cell there was failure.

The MB2-C specification does provide the capability for the BM-SC to notify the GCS AS about the status of an MBMS bearer using the MBMS Bearer Status Indication Procedure, as described in section 5.3.5 of 3GPP TS 29.468 [15]. The indication is provided in the GCS-Notification-Request command. The command contains the following elements:

```
<GN-Request> ::= < Diameter Header: 8388663, REQ, PXY >
                 < Session-Id >
                 [ DRMP ]
```

```
{ Auth-Application-Id }
{ Auth-Session-State }
{ Origin-Host }
{ Origin-Realm }
{ Destination-Realm }
{ Destination-Host }
[ Origin-State-Id ]
*[ Proxy-Info ]
*[ Route-Record ]
[ TMGI-Expiry ]
*[ MBMS-Bearer-Event-Notification ]
[ Restart-Counter ]
*[ AVP ]
```

Using this command, it is possible to report the status of an MBMS bearer. Currently only bearer exception occurring in the BM-SC itself can be reported in this manner using AVP MBMS-Bearer-Event-Notification. This could be extended with more values and AVPs to signal also exceptions in the network.

A.1.2.5 On xMB-C interface

The xMB interface provides for “pull” and “push” notification procedures. The “pull” notification is used if and when the content provider wishes to acquire information. Using the “push” notification, the content provider can indicate its interest in receiving notifications from the BM-SC which then can post the notification to the content provider. Following notification types or levels are supported:

- Critical: When some events drastically prevent the proper delivery of content, such as when the network is down, the data ingestion is interrupted, BM-SC data delivery function stopped, etc.
- Warning: When the service can be partially delivered but quality is reduced. The reason can be that the service is partly down because the data bitrate is too high, the packet loss rate is too high, etc.
- Information: When the service is properly delivered but some interesting event occurred. The reason can be the presence of reporting information for the service, the correct transmission of the service announcement, etc.
- Session/Service: Service/Session related parameters, such as service/session started, service/session terminated, Content file send, file fetching error, etc.

With this procedure richer information can be sent from the BM-SC to the application. The Session/Service notification level could be used to deliver information that an MBMS bearer has been pre-empted at one or more eNodeB if it is extended with some new properties. Using a JSON structure like the array, a new property could be introduced to the Release 14 structure to reporting success level on a per SAI basis as shown in the following example:

“Message Class”: “Session”,

“Message Name”: “SessionStateChange”,

“Session State”: “Session Active”,

“Message Source”: “5”,

“Active SAI”: [“11001”, “11003”, “11004”] ,

“Inactive SAI”: [“11002”]

A.1.2.6 Limitations affecting the eMBMS application

An application (MCPTT, video viewing, file delivery, PWS) requires an eMBMS bearer to be established in the network to deliver the relevant content. To establish a bearer, the

application requests the necessary resources using session management procedures through the BM-SC, possibly using MB2-C or xMB.

The application is required to provide the location where the bearer will be needed. This is done by providing the Service Area using a list of MBMS Service Area IDs (SAI) and/or a Cell Identities list (CI).

As the BM-SC currently cannot provide feedback to the application in case the bearer is not started in one or more eNodeBs, the application does not know in which location the resulting service succeeds or fails.

This lack of information could impact the application in the following ways:

1. Difficulty in applying a Service Level Agreement between content provider and network provider because of difficulty in measuring success
2. Inaccurate service reporting for statistics & analysis
3. Possible unnecessary resource usage
4. No ability to dynamically provide an alternative channel
5. No ability to dynamically provide repeats

The last two issues can be overcome by always providing an alternative channel or always providing repeats.

For example, the MCPTT application can dynamically switch over from MBMS to unicast as 3GPP specified means for a UE to report the unavailability of the MBMS session to the MCPTT server.

eMBMS allows the UE to report on reception using reception reports but this method does have its limitations:

- Reception reports are not real-time
- Difficulty to correlate a reception report to the true cause of error

With the xMB interface, the reception reports can be delivered from the BM-SC to the content provider, this resolves the issue that prior to 3GPP Release 14, reception reports could only be delivered to one destination (BM-SC or application).

A.1.2.7 Limitations affecting the Operations Support Systems

MNOs rely on various systems to keep the network running and satisfy their customers.

Fault Management. Based on faults reported by equipment in real-time, operations personnel may provide corrective action, depending on the severity and impact.

Not starting an eMBMS session or pre-empting an eMBMS session is not likely classified as a fault by the eNodeB and as result the event will not be reported for Fault Management.

Since the BM-SC is not aware of the event, it cannot report anything.

Trouble Ticketing. Provides the capability to log an issue/trouble for investigation.

As the inability to transmit in an area impacts multiple users and is not likely classified as a technical fault in Fault Management as explained above, it could be logged as a trouble ticket instead.

To log the trouble ticket electronically, the BM-SC or application does need to provide sufficient information to help with the investigation and as discussed in previous sections, this is still lacking.

Performance Management. Used to monitor the performance of the network over periods of time, deviations can be observed which may have sometimes their root cause in technical issues or lack of capacity. If an eMBMS bearer was pre-empted there is likely a capacity issue.

This requires the eNodeB to log the pre-emption event for Performance Management (which would be vendor specific) as well as other usage events indicating the high load. Since the event is not known to the BM-SC, the BM-SC cannot log it. However, it would be beneficial for the BM-SC to log the event together with the context information of the eMBMS session (which application, location etc.) which is not available at the eNodeB.

If the eNodeB doesn't log any event, detecting capacity issues causing exceptions impacting eMBMS would become very difficult.

Fault and Performance Management as well as Trouble Ticketing are often used by MNOs to manage Service Level Agreements (SLA) with various parties:

- Equipment suppliers
- Partners (MVNO, content suppliers)
- End-customers (large organizations)

With lack of KPI monitoring of eMBMS sessions, it is difficult to comply with an SLA for services utilizing eMBMS.

A.1.3 Lack of feedback from RAN to the core network

A.1.3.1 On MBMS delivery mode

As shown in Figure 24, when the BM-SC wants to activate an MBMS session, a Session Start Request message is sent through the MBMS-GW, the MME to the MCE. If an MBMS Cell List is provided the MCE then decides on whether to use MBSFN or SC-PTM as MBMS delivery mode according to clause 15.1.1 of 3GPP TS 36.300 [9]. However, the specification does not provide a guideline for the MCE to make an appropriate selection; this is left to implementation.

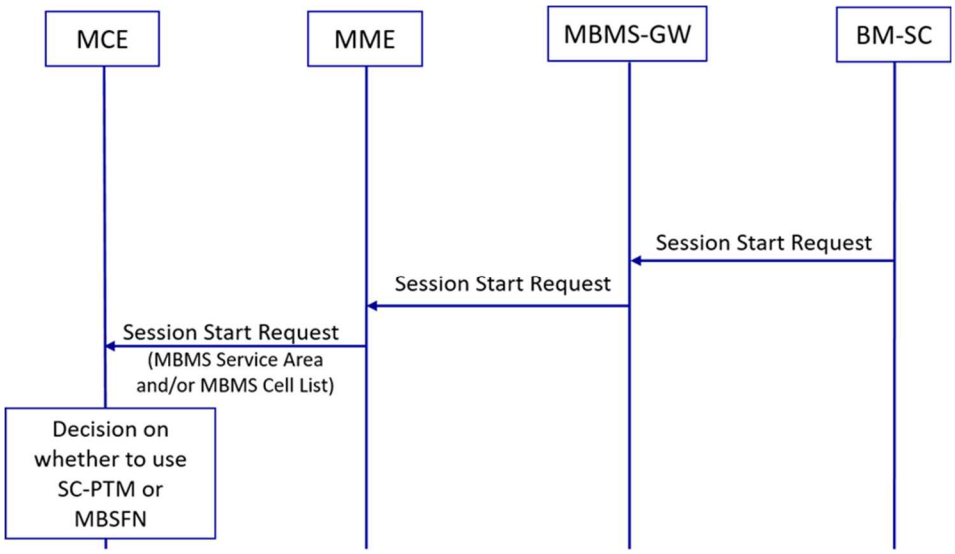


Figure 24: MBMS Start Procedure with MCE decision.

Clause 5.8 3 in 3GPP TR 23.780 [47] identifies a key issue on cohesive MBMS operation. One of the important points is that the core network does not know the MBMS mechanism selected by the MCE. Indeed, the information on the decision is not sent back to the core network to optimize the session parameters. One of the solutions

provided in 3GPP TR 23.780 has an approach where the core network indicates the preferred mechanism to the MCE. Then the MCE may take the decision based on this information and informs back to the core network as shown in Figure 25.

Another issue related to the cohesive MBMS operation comes from the UE capability which indicates the supported delivery modes (e.g. MBSFN, SC-PTM). It's desirable for the UE to send its MBMS capability information to the network so that the network store the MBMS capability from all UEs.

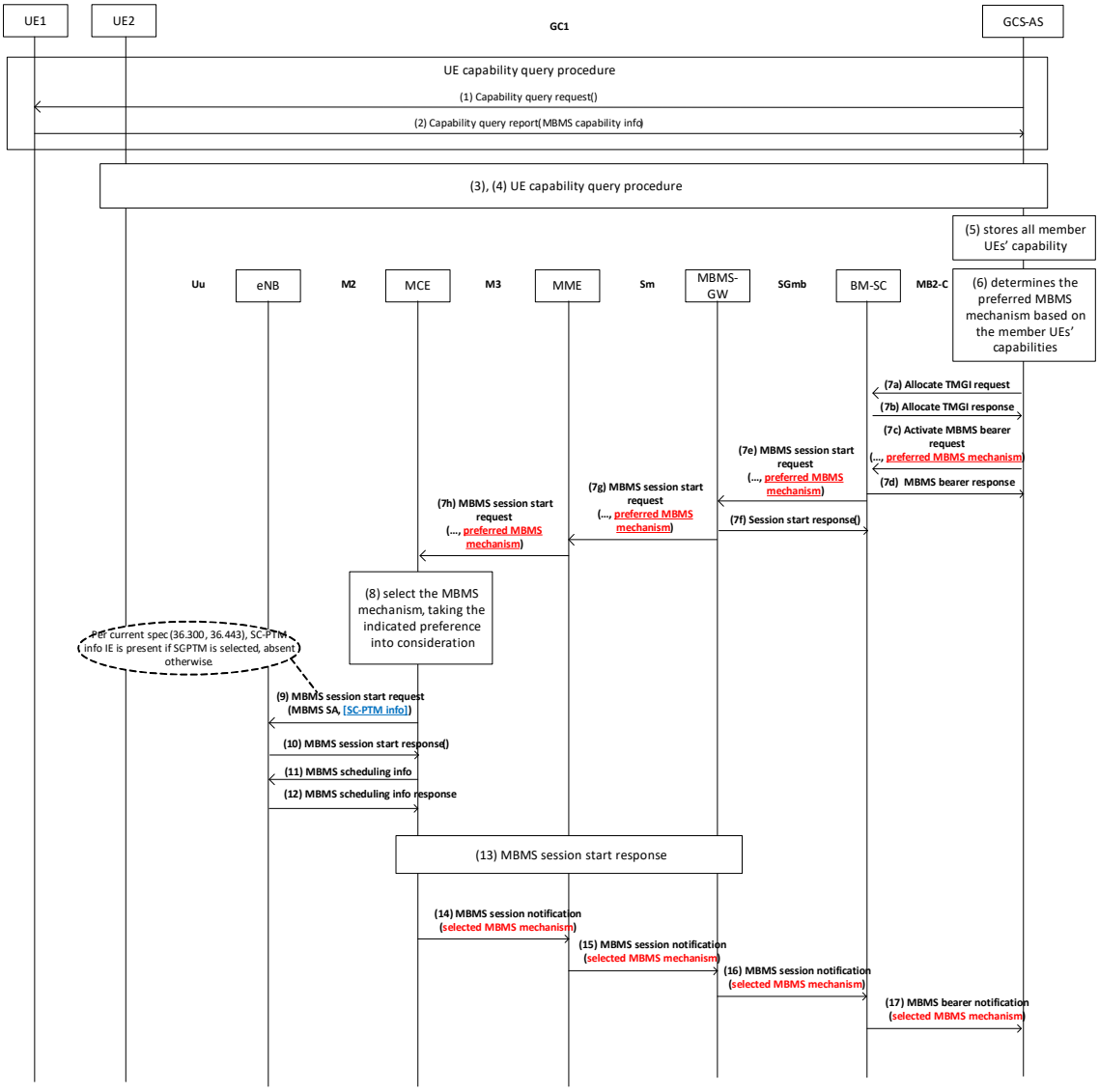


Figure 25: MBMS mechanism selection from 3GPP TR 23.780.

According to the TR conclusion, a decision on a solution should be evaluated in 3GPP SA2 due to its architectural nature. It will require the discussion with RAN working groups. At the time of writing this deliverable, it's considered as a limitation of eMBMS in 3GPP Release 14.

A.1.3.2 On MBMS scheduling parameters

The multi-cell MBSFN transmission in eMBMS architecture is not well suited for services whose QoS requirements vary over time (e.g. variable throughput and best-effort services). When the traffic characteristics of a service change, the content provider is

responsible for requesting an update on MBMS session parameters. There are two factors related to this issue.

Firstly, the MCE, which controls the MCCH service scheduling, resides in the control plane of eMBMS architecture. Hence, the MCE is not aware of the user plane traffic characteristics. The MCE determines MCCH scheduling information and possibly other configuration based on the information received in MBMS session management messages over M3 interface or in MBMS overload information from the eNodeB. The MBMS overload notification provides a binary indication whether PMCH is overloaded together with a list of active MBMS sessions. Both the information received in MBMS session management messages over M3 interface or the MBMS overload notification are insufficient in addressing the architectural drawback and allowing dynamic scheduling of multi-cell transmission.

Secondly, the structure of MCCH and MTCH with the corresponding procedures such as MCCH update were driven by the design of underlaying physical layer frame structure for MBSFN that relies on the semi-static segregation of MBSFN resources from unicast resources. The RRC (Radio Resource Control) signalling and the use of system information signalling of MBMS configuration inherit the semi-static nature of MBSFN. Altogether, MBSFN transmission is not suitable for dynamic scheduling. Moreover, the scheduling of MBSFN data transmission over the air is also impacted by the synchronization of user data through the core network, i.e. the use of synchronization sequences (SYNC protocol) between BM-SC and eNodeB.

Therefore, 5G-Xcast project aims at flexible session management and dynamic scheduling where the MBMS parameters can be dynamically changed during the MBMS session lifetime. This work is addressed in the deliverable D4.3 “Session control and management”.

However, for terrestrial broadcast over a large area up to nation-wide it has to be assumed that the QoS requirements are stable over time (e.g. fixed throughput and always-on transmission). In this specific case the requirements of the traffic characteristics and the QoS-targets have to be defined by the content provider in advance and the BM-SC has to configure the subsequent network elements (MBMS-GW, MME, eNodeB, etc.) accordingly.

A.1.3.3 On the SYNC protocol

3GPP TS 25.446 [11] defines the MBMS synchronization protocol (also called SYNC) that runs between the BM-SC and the eNodeB. The SYNC protocol is designed to ensure the synchronized transfer of MBMS user data from different cells to the UEs as shown in Figure 26. In addition, the SYNC protocol allows to detect lost or corrupted packets at the eNodeB. The specification in [11] specifies the eNodeB’s behaviour when detecting packet loss. In Figure 27, when detecting one lost packet and its size, the eNodeB fills the padding bit accordingly. If there are multiple consecutive lost packets, the eNodeB will stop the transmission until the next synchronization sequence.

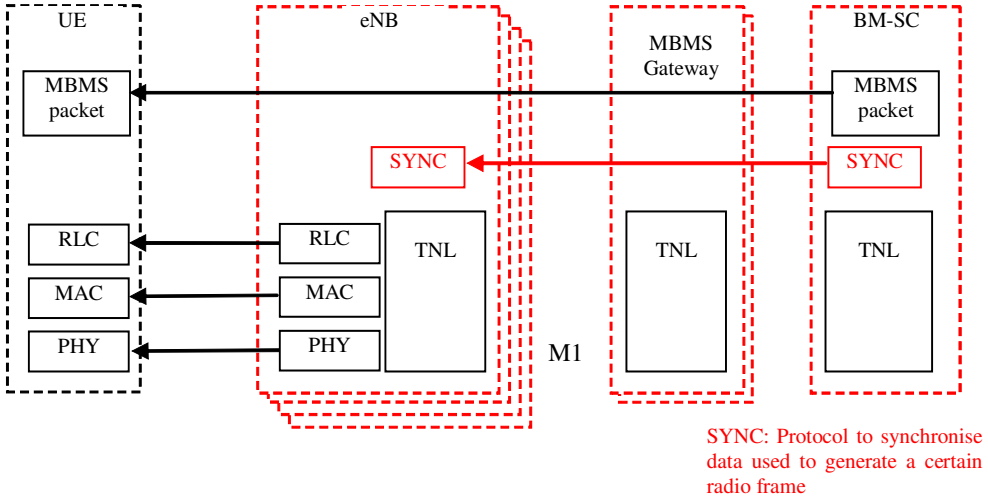


Figure 26: Overall u-plane architecture of the MBMS content synchronization.

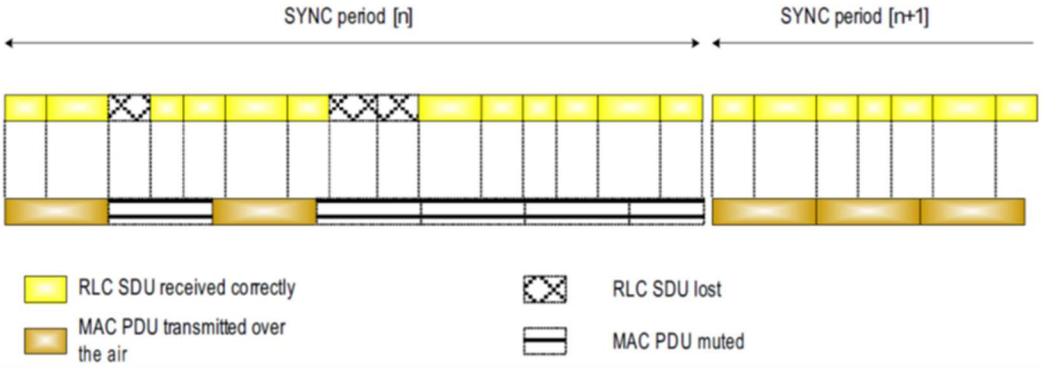


Figure 27. eNodeB behaviours when detecting lost packets.

Upon detection of lost packets, there is no indication from the eNodeB to inform the BM-SC on lost packets. The packet loss detected at the eNodeB can happen anywhere between the BM-SC and the eNodeB (i.e. on M1 or SGi-mb interfaces). It's desirable to monitor the KPI on the transport network with more details. It's important to separate the feedback mechanism from the eNodeB to the BM-SC in case of packet loss and the actions by the BM-SC upon reception of the feedback message.

The feedback on packet loss may help the BM-SC to monitor, troubleshoot if any major problems occur. From UE perspective, it cannot know where the loss is originating from (radio interface or connection between BM-SC and eNodeB), Reception Reports cannot help to indicate the root cause.

The lack of actions by the BM-SC may be explained by the fact that when the retransmitted packets arrive at the eNodeB, these packets may be outdated due to the duration of the synchronization sequence, especially when the loss is detected at the end of the synchronization sequence. Indeed, it is well known that retransmission of lost packets takes at least one round trip time (RTT) (Figure 28). It should be noted that the feedback from the eNodeB to the BM-SC doesn't exist today. The latency analysis should also take into account the transit and processing time at the MBMS-GW. In addition, the connection between the eNodeB and the BM-SC is more reliable than the radio interface between the eNodeB and UEs. Therefore, the packet loss in SYNC messages occurs less frequently. Furthermore, eMBMS also employs AL-FEC which

helps to recover the missing packets found by the UEs. This mechanism is able to mitigate the impacts of packet loss in SYNC protocol.

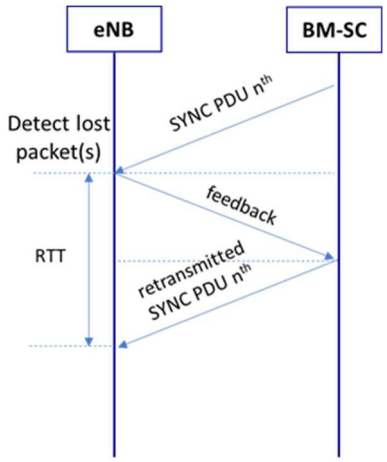


Figure 28. Latency analysis if eNodeB sends feedback when detecting lost packets.

To solve the issue, a feedback message for every SYNC sequence whether there are packet losses could be sent from all the eNodeBs within a MBSFN Area. It is then left to the BM-SC to have appropriate action depending on the situation.

SC-PTM does not need SYNC protocol since this delivery mode does not require the coordination between multiple cells. At the writing time of this deliverable, 3GPP decide to keep SYNC for SC-PTM. The reason is that neither the core network nor the content provider should need to know how RAN decides to transmit the MBMS content. Keeping the interface between the RAN and the core network the same, RAN operation is transparent to the rest of the network. However, there may be an issue with UE capabilities when a target UE does not support both transmission modes of which MCE is not aware.

A.1.4 Lack of efficient mechanism to trigger MBMS reception in the UE

In an LTE environment, the initiative to consume MBMS content relies on the UE. The UE has somehow become aware that the content of interest is going to be broadcasted or is being broadcasted and then takes the initiative to start up the application that shall fetch and display that content. MBMS user service can be advertised to UEs using the MBMS service announcement mechanism [10] which provides the information such as parameters required for service activation (e.g. IP multicast addresses, starting time etc.). The application on the UE then requests the MBMS client to receive a service which is characterized by a TMGI (Temporary Mobile Group Identity). In E-UTRAN, the modem (UE) then searches for the service in cells on frequencies indicated in the user service description file or frequencies indicated in SIB15 (System Information Block type15) for the TMGI. The UE acquires SIB2, SIB13, MCCH, SIB20 and SC-MCCH to determine whether the service is being transmitted. MCCH or SC-MCCH and possibly SIBs are updated when MBMS services is started in E-UTRAN. UE shall attempt to receive MCCH change notification and SC-MCCH change notification in each modification period. The change notification triggers UEs to reacquire MCCH or SC-MCCH and if the MBMS service of interest is included in the new information on MCCH or SC-MCCH then this is also the trigger for the UE to configure radio bearers for the service and receive the service.

In case of public warning (PW) messages a user is not aware when warning messages are broadcasted, because PW messages are not scheduled in advance, they are by nature unexpected. In order to receive a PW message once it is broadcasted, the modem

has to continuously monitor MCCH and SC-MCCH change notification, update MCCH and SC-MCCH information when the information changes, and search for the TMGI associated with PW Service within the new information. The MCCH and SC-MCCH modification periods can be set to wide range of values from 10's of milliseconds to a second for MCCH or even minutes for SC-MCCH. The MCCH and SC-MCCH modification periods shall be set in accordance with the service's latency requirement. When multiple services with diverse latency requirements are provided in the cell, the MCCH and SC-MCCH modification period configuration is driven by the most constrained requirement. This causes UE interested only in services with less constrained requirement to receive MCCH or SC-MCCH more frequently than necessary which may lead to unacceptable battery consumption depending how diverse the latency requirements are. This may be a significant issue for a PW message that is hopefully never broadcast.

Another challenge is that a TMGI value for PW is currently not standardized. If roaming of PW service is to be supported then the UE must be configured with MBMS service description for any network it can roam in or, alternatively, the MBMS service description could provide a list of alternative TMGIs if other parameters of the service are the same. In both cases, PW service requires coordination of MBMS configuration between MNOs. To potentially solve the issue, the TMGI for PW services could be the same in all networks both national and international roaming would be supported.

It's noted that eMBMS is not optimised for the support of services with various latency requirements which is a prominent issue in PW use cases due to the limitation described above. Section 4.2 explains further why eMBMS can be a solution for PW use cases using multimedia content (e.g. audio, video).

A.1.5 Service continuity when moving between MBSFN areas

When a UE moves from one MBSFN area to another MBSFN area (possibly in the same Service Area) then the UE uses information from USD or information received in SIB15 to prioritize MBSFN-areas on the indicated frequencies in the "cell" reselection process. When searching for a suitable MBSFN-area on the indicated frequencies, the UE should also attempt to acquire the information about MBMS services being provided in the neighbouring MBSFN-area via acquisition of SIB13 and MCCH prior to the reselection. This search process takes time and packet loss may occur in certain scenarios. For example, if the neighbouring MBSFN areas are using the same resources in the time and frequency domain, then the UE may not be able to acquire MBMS information from the neighbouring cell/MBSFN area due to the interference. 3GPP TR 23.780 [47] explains that for mission critical services the UE will temporarily switch to unicast and provides solutions to prevent this.

One solution documented in 3GPP TR 23.780 [47] is that MBSFN areas overlap, so the UE can already acquire the configuration information of the new MBSFN area when it is still in coverage of the old MBSFN area which it is about to leave. This mechanism requires new procedures in the UE to acquire scheduling information in overlapping MBSFN areas.

Another solution is that the UE performs cell reselection before coverage in the MBSFN area becomes a problem, which can be done when cells at the border of the MBSFN area also transmit scheduling information of the neighbour MBSFN area. Such a solution has RAN impact. Border cells in MBSFN areas shall transmit scheduling information of neighbour MBSFN areas. SC-PTM adopted the solution to some extend by including a list of neighbour cells providing MBMS services via SC-MRB (Single Cell MBMS Point to Multipoint Radio Bearer) in SC-MTCH configuration.

A.2 Suggestions for improvements

As the 5G architecture is different from the LTE, the following suggestions for improvements are applied in LTE. However, the following suggestions are taken into account in the design of 5G architecture having PTM enabled.

A.2.1 The role of the MME in the eMBMS architecture

As shown in Figure 1, the session control messages are sent by the MBMS-GW to the MME over the Sm reference point and M3 is the reference point for the session control messages between MME and MCE.

The MBMS-GW sends session control messages (i.e. MBMS-Session-Start, MBMS-Session-Update and MBMS-Session-Stop) over the Sm reference point to the MME and these messages contain the MBMS Service Area where the message needs to be transmitted and optionally an MBMS Cell List which allows the MCE to choose the SC-PTM method [17].

The MME forwards these session control messages over the M3 reference point to the MCEs that serve the Service Area [20].

The MME became aware of the Service Areas that the MCE serves via the M3 Setup Request which the MCE has sent to the MME upon initialization of the MCE or via M3 Configuration Update request.

The BM-SC receives the Service Area list through OAM configuration. This can be avoided if the M3 reference point didn't terminate on the MME but instead on the MBMS-GW. The MBMS-GW could then forward the Service Area list to the BM-SC in a standardized way, or alternatively, store the Service Area list in a Network Repository which is accessible by the BM-SC.

The functionality that resides in the MME to forward MBMS session management messages is an isolated functionality, it doesn't have any interface to any other functionality of the MME and therefore doesn't have to reside in the MME.

A.2.2 The role of the MBMS-GW in the eMBMS architecture

As shown in Figure 1, the MBMS-GW multicasts the user plane data to the eNodeBs over an IP multicast capable network. The eNodeBs that have joined the multicast group have received the multicast IP address via session control (from MBMS-GW, via MME, MCE to the eNodeBs that serve the Service Areas indicated in the MBMS session control message).

In deployments where MNOs share their RAN, but each MNO has its own core network, the IP multicast capable network may not lie in the same IP subnet. In such an environment, the MBMS-GW serves as gateway that separates IP subnets.

If there are two MBMS-GWs, which each act as a gateway into the RAN of an operator then the interfaces between BM-SC and each GW are SGmb and SGi-mb and the RANs are "behind" the gateways. If there is a single MBMS-GW only (or no GW at all) residing in the network of the operator that deploys the BM-SC then this gateway also interfaces with the entire RAN (IP backbone) of the sharing operator which may not be desired.

B Annex

B.1 Localized MBMS for V2X

3GPP has completed the specification of CUPS (Control User Plane Separation) in Release 14. The CUPS specification provides the architecture enhancements for the separation of control and user plane functionalities in the Evolved Packet Core's S-GW, P-GW and TDF. The separation enables flexible network deployment and operation (Figure 29).

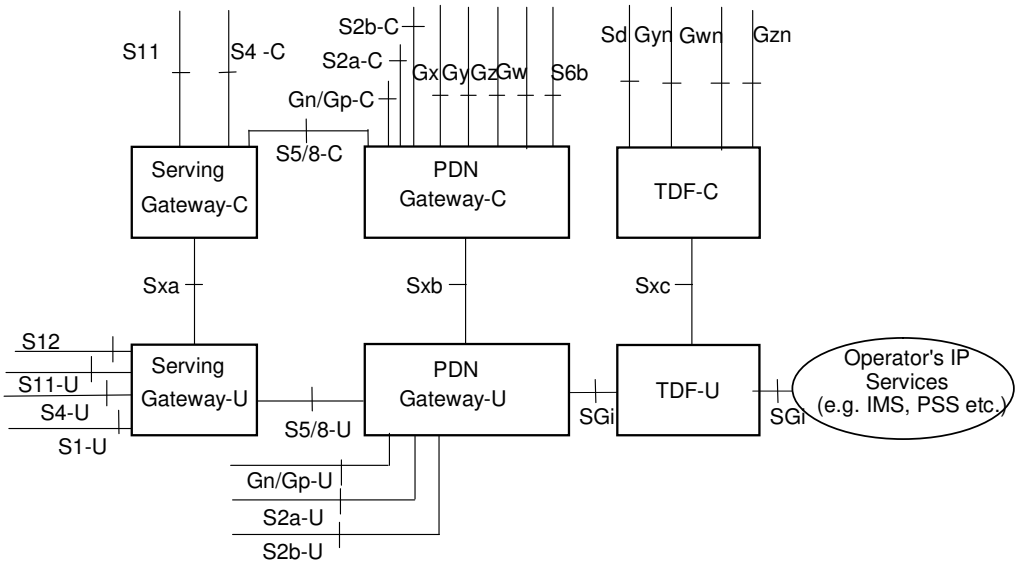


Figure 29. Architecture reference model with CUPS.

3GPP TS 23.285 [43] describes two options of localized MBMS deployment for the MBMS delivery of downlink V2X messages to vehicles in a V2X system. The localized MBMS deployment addresses the backhaul delay between the BM-SC and the eNodeB which is not negligible and can be achieved by either of the following deployment options:

1. To move the MBMS CN functions (e.g. BM-SC, MBMS-GW) closer to the eNodeB (Figure 30).
2. To move the user plane of MBMS CN functions (BM-SC, MBMS-GW) closer to the eNodeB (Figure 31).

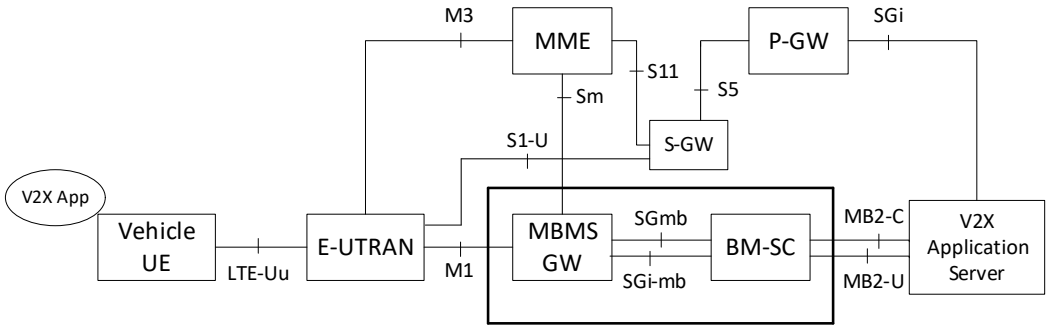


Figure 30. Localized MBMS CN functions (option 1).

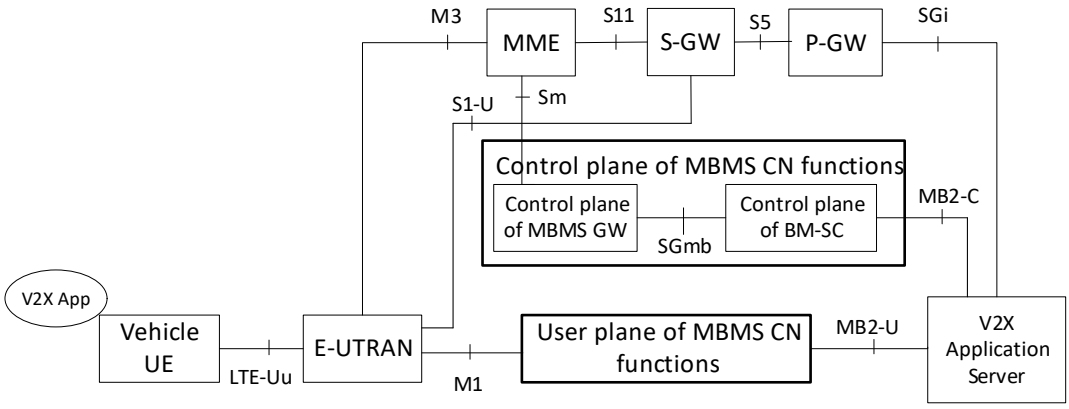


Figure 31. Localized user plane of MBMS CN functions (option 2).

The localized MBMS option 2 introduced the control and user plane separation for eMBMS similarly to the CUPS architecture. This separation is the key design principle of 5G system also followed by the 5G-Xcast project.