Satellite and Terrestrial Network for 5G

D3.2
Integrated SaT5G Detailed Network Architecture

<table>
<thead>
<tr>
<th>Topic</th>
<th>H2020-ICT-07-2017</th>
</tr>
</thead>
<tbody>
<tr>
<td>Project Title</td>
<td>Satellite and Terrestrial Network for 5G</td>
</tr>
<tr>
<td>Project Number</td>
<td>761413</td>
</tr>
<tr>
<td>Project Acronym</td>
<td>SaT5G</td>
</tr>
<tr>
<td>Contractual Delivery Date</td>
<td>M14 Final</td>
</tr>
<tr>
<td>Actual Delivery Date</td>
<td>17 December 2018</td>
</tr>
<tr>
<td>Contributing WP</td>
<td>WP3.2, WP3.3</td>
</tr>
<tr>
<td>Project Start Date</td>
<td>June 2017</td>
</tr>
<tr>
<td>Project Duration</td>
<td>30 months</td>
</tr>
<tr>
<td>Dissemination Level</td>
<td>Public</td>
</tr>
<tr>
<td>Editor</td>
<td>Part A: ADS</td>
</tr>
<tr>
<td></td>
<td>Part B: BPK</td>
</tr>
<tr>
<td>Contributors</td>
<td>Part A: ADS, AVA, BT, GLT, i2CAT, SES, UoS, ZII</td>
</tr>
<tr>
<td></td>
<td>Part B: ADS, AVA, BPK, BT, GLT, i2Cat, iDirect, QUO, SES, TNO, UOS</td>
</tr>
</tbody>
</table>
Executive summary

SaT5G aims to deliver the seamless integration of satellite into 5G networks to ensure ubiquitous 5G access everywhere.

This document describes the accomplishments of two work packages:

- **WP3.2 – Backhaul Architectures** whose purpose is to provide a detailed analysis of the topics which need to be taken into account for the design of future 5G satellite-terrestrial backhaul. The integrated satellite-terrestrial reference network and the backhaul implementation options defined in D3.1 – Integrated SaT5G general network architecture, as well as the SaT5G use cases and scenarios defined in D2.1 – Satellite Reference Use Cases and Scenarios for eMBB and the ongoing 5G standardisation, are key inputs for this document. These aspects are addressed in the part A of the document;

- **WP3.3 – Caching and Multicast Architectures** whose purpose is to design detailed network architecture that focus on the delivery of content to the edge and multicast solutions leveraging on the ubiquity and broadcast/multicast capabilities of satellite systems. These aspects are addressed in the part B of the document.

Part A: Backhaul Architectures

As shown in D3.1, it has been determined that satellite-terrestrial integration will require advanced functionalities in the framework of each network, or at the interface between the two networks, to provide a high-value integrated network. Analysis done in the framework of this document has led to the identification and preliminary analysis of several high-level functionalities for satellite and terrestrial integration:

- Registration, connection and roaming management of the gNB or the relay node;
- Management of the end to end quality of service;
- Mobility management and handover in the satellite systems;
- Multilink management;
- Support of network slicing and requirement for network function virtualisation.

The consideration of the backhaul implementation options investigated in D3.1 – Integrated SaT5G General Network Architecture have let to the conclusion that in general:

- For the backhaul implementation options based on the relay node, a single network management system, typically the MNO 3GPP NMS, is involved, and defined or under definition 5G system specifications applicable to the UE can be adopted and adapted if necessary for the implementation of the identified functionalities at the relay node level;
- For the backhaul implementation options based on transport network, the 3GPP NMS and the transport network NMS will have to work in a close coordination in order to properly operate the identified high level functionalities, by the insurance of sufficient exposure of TN status to the terrestrial system, the support to resource allocation, the quality of service adaptation and the support of network slicing.

These aspects are summarised in the Table 0-1 below, with the implementation requirements of the identified high level functionalities, for the backhaul implementations options, based on RN in one hand and based on transport network in the other hand.

The flexibility at different levels is identified as a key requirement for an efficient 5G satellite and terrestrial integration. Such flexibility is enabled by the adoption of SDN/NFV paradigms in the frame of NTN domain, but also by the introduction of generic and flexible components. We can introduce in this context the concept of “generic NTN terminal”, which would be able to connect to any NTN system (GEO from different SNOs, MEO, HAPS, etc.), and which can be implemented either by the combination of various terminals for each NTN system, or by a flexible and powerful terminal being able to switch between NTN systems or being simultaneously connected to several NTN systems. We can also introduce the concept of “generic gNB”, which would be able to connect to any 3GPP 5G Core from any MNO without requiring a specific attachment or belonging to a particular MNO.
Table 0-1: Requirement for the implementation of high-level features for 5G satellite and terrestrial integration

<table>
<thead>
<tr>
<th>High-level features applied to integrated satellite and terrestrial 5G network</th>
<th>Satellite backhaul implementation options based on relay node</th>
<th>Satellite backhaul implementation options based on Transport network</th>
</tr>
</thead>
<tbody>
<tr>
<td>Registration, connection and roaming management</td>
<td>The RN reuses the registration and connection procedure as defined for the UE in 3GPP</td>
<td>The UE registration, connection and roaming is not affected</td>
</tr>
<tr>
<td></td>
<td>RN acts as a proxy for the UE registration and connection to the 5G core network</td>
<td>The “generic gNB” performs the registration to the 3GPP core that will serve</td>
</tr>
<tr>
<td>Resource allocation and orchestration</td>
<td>Resources manage by 3GPP NMS, including satellite resources</td>
<td>Coordination is necessary between the 3GPP NMS and the NTN NMS for the allocation of the NTN resources</td>
</tr>
<tr>
<td>Management of the end to end quality of service</td>
<td>Mapping from 5QI (3GPP identifier of the quality of service) to sat resources</td>
<td>Mapping from 5QI (3GPP identifier of the quality of service) to sat CoS (NTN class of service identifier) and from sat CoS to sat resources</td>
</tr>
<tr>
<td>Mobility management and handover in the satellite systems</td>
<td>Mobility and handover, including the satellite beam handover are managed by the 3GPP NMS</td>
<td>Coordination is necessary between the NTN NMS and the 3GPP NMS. The mobility management of the gNB is managed by the 3GPP NMS whereas satellite beam handover remain managed by the NTN NMS</td>
</tr>
<tr>
<td>Multilink management</td>
<td>The AT3SF specification for traffic steering, splitting and switching for the UE is adapted for the RN to enable multi-link-based backhaul implementation</td>
<td>Specified Hybrid Multiplay Functions are implemented between the gNB and the NTN terminal and between NTN GW and 5G core, the objective being to perform traffic steering, splitting and switching at backhaul level</td>
</tr>
<tr>
<td>Support of network slicing and requirement for network function virtualisation</td>
<td>E2E slicing supported and managed by 3GPP management, assuming that NTN domain management is integrated in 3GPP NMS</td>
<td>Management of the NTN domain needs to be in coordination with 3GPP NMS. The adoption of SDN/NVF paradigms in the NTN domain is recommended to enable the efficient support of network slicing</td>
</tr>
</tbody>
</table>
Part B: Caching and Multicast Architectures

The main challenge addressed in the part B consists of complementing the existing mobile network with satellite connectivity in order to bring HD and 4K content to areas with poor terrestrial connectivity and on moving platforms such as planes or vessels in order to improve end user QoE. The end-to-end latency also has to be minimised, especially for live events that generate a huge audience, like sports events.

- **Rationale:**

Currently, serving mobile devices on-the-go is performed from a central Point of Presence (POP) controlled by the network operator. Delivery to mobile devices does not usually benefit from the CDN concept that consists of caching the most popular content at the edge and streaming this from a location closer to end users. In current mobile networks (i.e. LTE networks), there are no physical locations at the base-stations where the content can be cached. It is always a given POP that is elected to stream all the content to mobile users. Even if servers can be added in this POP to increase its efficiency in terms of bandwidth output, the central site connectivity can become the bottleneck. Leveraging new hosting locations such as base stations would lead to a higher granularity of POPs and the possibility of streaming content from a location closer to end users, thus improving its delivery. These POPs need to be provisioned with the content to stream, with a certain level of elasticity to cache in each location only the most popular content.

- **Solutions:**

Two main topics have been investigated for an efficient content delivery over the satellite.

- **Asset caching for video delivery:** At first, we addressed the asset caching, focusing on video delivery since it’s the most popular content both in consumption and current traffic over the network [1]. The solution envisaged could be easily extended to other types of assets (Webpages, VNF software update repository, etc.). We designed a generic architecture reusing satellite multicast capabilities to cache popular (or expected to be popular) assets to the edge. Within the Core Network, a dedicated AF for caching is deployed. The AF is in charge of steering traffic to the Edge Network via the satellite link so that it is served locally from the cache server (located in Local DN) to the end users. This AF is also in charge of orchestrating the caching by multicasting to the edge the popular assets in a carousel. Popular assets stored in the carousel are pushed to the edge one after the other until the end is reached; afterwards we restart from the beginning and so on. The Local cache joins this multicast and stores locally the assets chunks. This implementation solution addresses the offline caching of popular assets, therefore not being triggered by any end user request. The asset popularity is computed on the fly by an analytics server. For more dynamism, another solution based on prefetching of segments has also been studied;

- **Optimising live OTT video delivery:** Secondly, we investigated how to optimize the delivery of over-the-top (OTT) video live channels. As a basis, we reused the work done on the Multicast Adaptive Bitrate (mABR). The basic concept behind mABR, is to send in multicast the live channel/stream over the backbone (the satellite backhaul in our case) then convert it back to unicast in the edge of the network. To achieve this, we defined a high-level architecture similar to the one presented for offline content caching.

- **Implementation considerations:**

The development of prototypes for these architecture proposals will be done during the project with a major milestone in late 2019 where demonstrations will take place. Simulation and implementation of the architecture will be done in the scope of WP4.6. The developed functionalities will afterwards be integrated, validated and demonstrated in two testbeds: at the University of Surrey (5GIC Testbed) and in Zodiac Aerospace premises (ZII Testbed).

- In 5GIC testbed, we focus on the demonstration of live channel delivery using the mABR solution and dynamic prefetching of segments for still content. These demonstrations will show a massive reduction on satellite resource usage as well as an improvement of overall QoE: use of higher layers, reduction of latency.
- ZII testbed will hold a demonstration of the offline caching. An aircraft environment is simulated; this aircraft has a local caching server on-board which is updated regularly by a multicast carousel containing the most popular assets. The purpose here is to demonstrate
the reduction of the required satellite bandwidth usage when popular assets are played from
the aircraft.

In these two demonstrations, a dedicated library will be integrated in the UEs aiming at retrieving QoE
indicators such as latency, rebuffing and stalls. These indicators will be put in relation with the KPIs
defined in WP2.

- Expected impacts:

The proposed solutions are expected to impact the way mobile network operators will invest in their
network in the 5G context. The availability of a Multi Access Edge Computing (MEC) environment will
allow deploying in a dynamic manner every local caching server that can be provisioned with popular
content through a satellite contribution link. The services addressed can be both for live and on-
demand content. From the user point of view, a global increase in terms of QoE is expected: fewer
rebuffering, higher bitrates and reduced end-to-end latency for live content. From the operator point of
view, caching popular assets to the edge and using multicast for live delivery via satellite will bring a
major reduction of network resources usage allowing the deployment of new high demanding services
like VR or 4K video streaming.

Next steps

The analysis performed throughout the parts A and B of the document have led to some valuable
recommendation of 5G standardisation, and to the definition or amendments of research pillars for the
WP4 – Research to Prototype Development.

The requirement for standardisation work includes:

- The adaption of UE specification for the backhaul implementation options based on relay
  node, in order to perform: the management and registration, the mobility management of the
  RN, the multilink management using an adaptation of the ATSSS and the resource allocations
to support end to end network slicing;

- The introduction of the concept of generic gNB in 3GPP to in order to allow the introduction of
  a third-party that can deploy a gNB and through a broker, can trade the 5G Core to serve that
gNB. This particular situation has been investigated for the specification of the detailed
network to address the SaT5G scenario 4a: the airline company in this case has been
identified as the third-party;

- Further specification of the coordination between the 3GPP NMS and the TN NMS, in
  particular for the exposure of the TN status, the TN resource allocation, and the support of
  network slicing;

- Quality of service adaptation and TN QoS awareness in order to provide feedbacks to the
  status of the TN to 3GPP NMS regarding the support of a specific required QoS;

- The introduction of a dedicated Caching Application Function in charge of applying redirection
  policies to the local caches for the optimization of the end-to-end delivery;

- The addition of local caches within the Edge Network filled by a satellite multicast transport for
  offline caching and/or by a prefetching mean for online caching.

In other SaT5G work packages, further analysis will be performed on the following topics:

- NFV, network slicing requirements and management an orchestration can be further
  investigated in WP4.1 – Implementing 5G SDN and NFV in Satellite Networks and WP4.2 –
  Integrated Network Management and Orchestration;

- Multilink management in WP4.3 – Multilink and heterogeneous transport: the work may
  include development of the adaptation of ATSSS for the RN (RN-ATSSS) and analysis of the
  distribution protocol (MP-TCP and MP-QUIC);

- Development of the QoS and definition of the mapping table in WP4.4 – Harmonisation of
  Satcom with 5G Control and User Planes;

- The proposed caching and multicast architectures features and architectures investigated
  throughout the part B will be further analysed and implemented in WP4.6 – Caching &
  multicast for optimised content & NFV distribution and demonstrated in WP5.3 – 5G Cellular
  and Home “Plug’n Play” and WP5.4 – 5G Moving Platform.
Satellite and Terrestrial Network for 5G

D3.2
Integrated SaT5G Detailed Network Architecture

Part A
Backhaul Architectures

<table>
<thead>
<tr>
<th>Topic</th>
<th>H2020-ICT-07-2017</th>
</tr>
</thead>
<tbody>
<tr>
<td>Project Title</td>
<td>Satellite and Terrestrial Network for 5G</td>
</tr>
<tr>
<td>Project Number</td>
<td>761413</td>
</tr>
<tr>
<td>Project Acronym</td>
<td>SaT5G</td>
</tr>
<tr>
<td>Contractual Delivery Date</td>
<td>M14 Final</td>
</tr>
<tr>
<td>Actual Delivery Date</td>
<td>17 December 2018</td>
</tr>
<tr>
<td>Contributing WP</td>
<td>WP3.2</td>
</tr>
<tr>
<td>Project Start Date</td>
<td>June 2017</td>
</tr>
<tr>
<td>Project Duration</td>
<td>30 months</td>
</tr>
<tr>
<td>Dissemination Level</td>
<td>Public</td>
</tr>
<tr>
<td>Editor</td>
<td>ADS</td>
</tr>
<tr>
<td>Contributors</td>
<td>AVA, UoS, SES, BT, ZII, GLT, i2CAT</td>
</tr>
<tr>
<td>Version</td>
<td>Date</td>
</tr>
<tr>
<td>---------</td>
<td>------------</td>
</tr>
<tr>
<td>00.01</td>
<td>22/11/2017</td>
</tr>
<tr>
<td>00.02</td>
<td>01/05/2028</td>
</tr>
<tr>
<td>&gt;&gt;</td>
<td>15/07/2018</td>
</tr>
<tr>
<td>&gt;&gt;</td>
<td>17/07/2018</td>
</tr>
<tr>
<td>00.19</td>
<td>25/10/2018</td>
</tr>
<tr>
<td>00.23</td>
<td>08/11/2018</td>
</tr>
<tr>
<td>&gt;&gt;</td>
<td>08/11/2018</td>
</tr>
<tr>
<td>&gt;&gt;</td>
<td>12/11/2018</td>
</tr>
<tr>
<td>0.27</td>
<td>15/11/2018</td>
</tr>
<tr>
<td>0.28</td>
<td>16/11/2018</td>
</tr>
<tr>
<td>&gt;&gt;</td>
<td>19/11/2018</td>
</tr>
<tr>
<td>&gt;&gt;</td>
<td>21/11/2018</td>
</tr>
<tr>
<td>&gt;&gt;</td>
<td>27/11/2018</td>
</tr>
<tr>
<td>&gt;&gt;</td>
<td>28/11/2018</td>
</tr>
<tr>
<td>0.31</td>
<td>28/11/2018</td>
</tr>
<tr>
<td>1.00</td>
<td>05/12/2018</td>
</tr>
<tr>
<td>1.50</td>
<td>13/12/2018</td>
</tr>
<tr>
<td>2.00</td>
<td>17/12/2018</td>
</tr>
<tr>
<td>Name</td>
<td>Organisation</td>
</tr>
<tr>
<td>---------------------</td>
<td>--------------</td>
</tr>
<tr>
<td>Boris Tiomela Jou</td>
<td>ADS</td>
</tr>
<tr>
<td>Paul Foulon</td>
<td>ADS</td>
</tr>
<tr>
<td>Simon Watts</td>
<td>AVA</td>
</tr>
<tr>
<td>Leonardo Goratti</td>
<td>ZII</td>
</tr>
<tr>
<td>Christos Politis</td>
<td>SES</td>
</tr>
<tr>
<td>Patrice Bertin</td>
<td>ADS</td>
</tr>
<tr>
<td>Pouria Sayyad Khodashenas</td>
<td>i2cat</td>
</tr>
<tr>
<td>Thierry Masson</td>
<td>EKY</td>
</tr>
<tr>
<td>Oriol Vidal</td>
<td>ADS</td>
</tr>
<tr>
<td>Keith Briggs</td>
<td>BT</td>
</tr>
<tr>
<td>Konstantinos Liolis</td>
<td>SES</td>
</tr>
<tr>
<td>Raz Ben-Haim</td>
<td>GLT</td>
</tr>
<tr>
<td>Barry Evans</td>
<td>UoS</td>
</tr>
</tbody>
</table>
# Table of contents

Executive summary ................................................................................................................................. 2  
Part A: Backhaul Architectures ............................................................................................................ 2  
Part B: Caching and Multicast Architectures ....................................................................................... 4  
Next steps ............................................................................................................................................ 5  
List of figures ......................................................................................................................................... 11  
List of tables .......................................................................................................................................... 12  
List of acronyms .................................................................................................................................... 13  
1 Introduction .................................................................................................................................... 16  
   1.1 Document context ................................................................................................................. 16  
   1.2 Organisation of the document ............................................................................................... 17  
2 Generic backhauling architecture and implementation options ..................................................... 18  
   2.1 Satellite backhauling state-of-the-art .................................................................................... 18  
      2.1.1 Backhaul via GEO satellites .............................................................................................. 18  
      2.1.2 Backhaul via non-GEO satellites ....................................................................................... 20  
      2.1.3 Satellite backhaul access techniques ............................................................................... 21  
   2.2 Requirements for 5G ready satellite systems ....................................................................... 23  
   2.3 SaT5G reference architecture ............................................................................................... 23  
   2.4 Satellite backhaul implementation options in 5G networks ................................................... 24  
      2.4.1 General concepts .............................................................................................................. 24  
      2.4.2 Relay node ........................................................................................................................ 25  
      2.4.3 Transport network ............................................................................................................. 25  
      2.4.4 Analysis of the backhaul implementation options ............................................................. 25  
3 High-level features applied to integrated satellite and terrestrial 5G network ............................... 27  
   3.1 Registration, connection and roaming management ............................................................ 27  
      3.1.1 Registration and connection management ....................................................................... 27  
      3.1.2 Generic gNB-satellite terminal .......................................................................................... 28  
   3.2 Resource allocation and orchestration .................................................................................. 29  
   3.3 Quality of service ................................................................................................................... 30  
      3.3.1 Legacy Sat QoS and requirement for evolution ................................................................. 30  
      3.3.2 Overview of QoS in 5G network overview .................................................................... 32  
      3.3.3 QoS management for relay node implementation options ................................................ 35  
      3.3.4 QoS management for Transport Network implementation options ................................... 36  
   3.4 (Radio) Access Network aspects ............................................................................................ 38  
      3.4.1 Mobility management in 3GPP Release 15 ...................................................................... 38  
      3.4.2 Handover in legacy satellite system .................................................................................. 40  
      3.4.3 Challenges of mobility management in SaT5G ................................................................. 41  
   3.5 Multilink management ........................................................................................................... 51  
      3.5.1 Protocols for traffic steering, splitting and switching distribution ........................................ 51  
      3.5.2 Relay-node implementation ............................................................................................... 53
3.5.3 Transport network implementation ................................................................. 54
3.6 Slicing and virtualisation ..................................................................................... 55
  3.6.1 Management of network slicing and slice definition criteria ......................... 55
  3.6.2 Network Function Virtualisation requirements for network slicing ............... 57
3.7 Features related to the implementation options .................................................. 57
  3.7.1 Transport network non-based on 3GPP system specifications ....................... 58
  3.7.2 Transport network based on 3GPP system specifications ............................... 59
  3.7.3 Relay node based on untrusted non-3GPP access ........................................... 59
  3.7.4 Relay node based on trusted non-3GPP access .............................................. 60
  3.7.5 Relay node based on 3GPP access ................................................................. 61
4 Network deployment analysis .................................................................................. 63
  4.1 Considered role model and backhaul implementation options ............................ 63
    4.1.1 Role model .................................................................................................... 63
    4.1.2 Considered backhaul implementation options .............................................. 66
  4.2 Network design for SaT5G scenario 2b .............................................................. 67
    4.2.1 Scenario requirements .................................................................................. 67
    4.2.2 Negotiation scheme between the actors for the SaT5G scenario 2b ............... 67
    4.2.3 Integrated backhaul architecture for SaT5G scenario 2b ............................... 67
  4.3 Network design for SaT5G scenario 4a .............................................................. 69
    4.3.1 Scenario requirements .................................................................................. 69
    4.3.2 Negotiation scheme between the actors for SaT5G scenario 4a ..................... 70
    4.3.3 Integrated backhaul architecture for SaT5G scenario 4a ............................... 70
5 Recommendations for standardisation and for WP4 research pillars .................... 73
  5.1 Recommendations for standardisation ............................................................... 73
  5.2 Recommendations for WP4 research pillars ...................................................... 74
6 Conclusion ............................................................................................................. 75
7 Bibliography .......................................................................................................... 78
List of figures

Figure 1-1: WP3 Strategy approach and SWP3.x interaction..................................................... 16
Figure 2-1: Evolving cellular use cases ....................................................................................... 19
Figure 2-2: Silat diagram showing LTE satellite backhaul and TCP acceleration....................... 19
Figure 2-3: Indirect diagram showing 3G cellular backhaul architecture ..................................... 20
Figure 2-4: MEO based backhaul offerings ................................................................................. 21
Figure 2-5: shared TDMA for variable traffic volumes, SCPC for steady traffic volumes [8] ............ 22
Figure 2-6: Number of sites required to justify TDM/TDMA vs. SCPC [9] ....................................... 22
Figure 2-7: SaT5G reference architecture ..................................................................................... 24
Figure 2-8: Backhaul implementation options identified by SaT5G ................................................ 24
Figure 2-9: SaT5G relay node approach ....................................................................................... 25
Figure 3-1: RN registration and UE registration forwarding.......................................................... 28
Figure 3-2: Generic NTN Terminal and gNB .................................................................................. 28
Figure 3-3: 5G QoS model overview ............................................................................................ 32
Figure 3-4: Classification, marking and mapping of the user plane traffic to access network resources .................................................................................................................. 33
Figure 3-5: QoS process from IP packets to DRB .......................................................................... 35
Figure 3-6: SDAP positioning illustration ..................................................................................... 36
Figure 3-7: QoS adaptation configuration for TN implementation options .................................... 37
Figure 3-8: SDAP positioning illustration ..................................................................................... 37
Figure 3-9: Intra AMF/UPF handover with Xn interface in place – [15], [16] ................................. 40
Figure 3-10: Legacy satellite terminal mobility management based on geographical location ........ 41
Figure 3-11: Example scenario for studying mobility management in case of an aircraft moving platform ........................................................................................................................................ 44
Figure 3-12: Single-gateway multi-beam satellite for aircraft moving platform case ....................... 45
Figure 3-13: Multi-gateway multi-beam satellite for aircraft moving platform case ......................... 46
Figure 3-14: Mobility management in 5G integrated satellite network in TN implementation option with single GW that does not require mobile network handover but satellite beam management .................................................. 47
Figure 3-15: Mobility management in 5G integrated satellite network in RN implementation option with single GW that does require handover in multiple gateways case with single CU and multi-DU .......... 48
Figure 3-16: Path switch selection in single GW single CU multi-DU IAB satellite with mobile relay... 48
Figure 3-17: Mobility management in 5G integrated satellite network with multiple GWs and CUs ....... 50
Figure 3-18: IAB node handover procedure in RN when changing IAB Donor node in multi-GWs, multi-beam 5G integrated satellite system and IAB architecture implements CU-DU functional split.. 50
Figure 3-19: Multi-satellite connectivity ....................................................................................... 51
Figure 3-20: Example MPTCP Usage Scenario in HMF implementation ........................................ 52
Figure 3-21: HTTP/2+TLS+TCP compared to QUIC .................................................................... 53
Figure 3-22: Comparison in High RTT, 2% loss environment ...................................................... 53
Figure 3-23: AT3S in a Relay Node implementation ....................................................................... 54
Figure 3-24: MPTCP implemented in HMF function .................................................................... 54
Figure 3-25: Cross-domain slice ................................................................................................. 55
Figure 3-26: Slicing management in relay node case ...................................................................... 56
Figure 3-27: Slicing management in transport network case ........................................................ 56
Figure 3-28: Transport network non-based on 3GPP system specifications .................................... 58
Figure 3-29: Transport network based on 3GPP system specifications ......................................... 59
Figure 3-30: Relay node based on untrusted non-3GPP access ....................................................... 60
Figure 3-31: Structure of TNAN [25] ........................................................................................... 60
Figure 3-32: Multi-TNAP/TNGF architecture [25] ....................................................................... 61
Figure 3-33: Relay node based on trusted non-3GPP access .......................................................... 61
Figure 3-34: Relay node based on 3GPP access ............................................................................ 62
Figure 4-1: Role model 2 (Several MNOs sharing Satellite resources) ........................................... 63
Figure 4-2: Broker entity layered view representation ..................................................................... 64
Figure 4-3: Example of workflow to negotiate satellite resources through a broker ....................... 66
Figure 4-4: Backhaul architecture for SaT5G scenario 2b with RN implementation option ............... 68
Figure 4-5: Backhaul architecture for SaT5G scenario 2b with TN implementation option ............... 69
Figure 4-6: Backhaul architecture for SaT5G scenario 4a with RN implementation option ............... 71
Figure 4-7: Backhaul architecture for SaT5G scenario 4a with TN implementation option ............... 72
List of tables

Table 0-1: Requirement for the implementation of high-level features for 5G satellite and terrestrial integration................................................................................................................................................ 3
Table 2-1: Key challenges for the SaT5G backhaul implementation options and associated timescale .............................................................................................................................................................. 26
Table 3-1: Satellite Service Classes and impact.......................................................................................................................................................................................... 31
Table 3-2: 5G QoS model acronyms .............................................................................................................................................................................................. 33
Table 3-3: Standardised 5QI to QoS characteristics mapping [14]............................................................................................................................................. 34
Table 3-4: Example of a 5QI to satellite resources mapping table.............................................................................................................................. 36
Table 3-5: Example of mapping table between 5QI and sat CoS............................................................................................................................................. 38
Table 3-6: Example of mapping table between Sat CoS and satellite resources............................................................................................................................................. 38
Table 3-7: Configurations for mobility management in SaT5G context............................................................................................................................................. 42
Table 4-1: SaT5G scenario 2b analysis.............................................................................................................................................................................................. 68
Table 4-2: SaT5G scenario 4a analysis.............................................................................................................................................................................................. 71
Table 6-1: Requirement for the implementation of high-level features for 5G satellite and terrestrial integration .............................................................................................................................................................. 75
### List of acronyms

<table>
<thead>
<tr>
<th>Acronym</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>3GPP</td>
<td>3rd Generation Partnership Project</td>
</tr>
<tr>
<td>5GCN</td>
<td>5G Core Network</td>
</tr>
<tr>
<td>5GTN</td>
<td>5G Transport Network</td>
</tr>
<tr>
<td>ACM</td>
<td>Adaptive Coding and Modulation</td>
</tr>
<tr>
<td>ADSL</td>
<td>Asymmetric Digital Subscriber Line</td>
</tr>
<tr>
<td>AF</td>
<td>Application Function</td>
</tr>
<tr>
<td>AKA</td>
<td>Authentication and Key Agreement</td>
</tr>
<tr>
<td>AMF</td>
<td>Access and Mobility Management function</td>
</tr>
<tr>
<td>API</td>
<td>Application Programming Interface</td>
</tr>
<tr>
<td>ARPF</td>
<td>Authentication Credential Repository and Processing Function</td>
</tr>
<tr>
<td>AS</td>
<td>Access Stratum</td>
</tr>
<tr>
<td>AT3SF</td>
<td>Access Traffic Steering, Splitting and Switching</td>
</tr>
<tr>
<td>AUSF</td>
<td>Authentication Server Function</td>
</tr>
<tr>
<td>BSS</td>
<td>Business Support Systems</td>
</tr>
<tr>
<td>BT3SF</td>
<td>Backhaul Traffic Steering, Splitting and Switching</td>
</tr>
<tr>
<td>CDN</td>
<td>Content Delivery Network</td>
</tr>
<tr>
<td>CIR</td>
<td>Committed Information Rate</td>
</tr>
<tr>
<td>CMF</td>
<td>Control and Monitoring Functions</td>
</tr>
<tr>
<td>CN</td>
<td>Core Network</td>
</tr>
<tr>
<td>CoS</td>
<td>Class of Service</td>
</tr>
<tr>
<td>DASH</td>
<td>Dynamic Adaptive Streaming over HTTP platform</td>
</tr>
<tr>
<td>DC</td>
<td>Data Centre</td>
</tr>
<tr>
<td>DN</td>
<td>Data Network</td>
</tr>
<tr>
<td>DNS</td>
<td>Domain Name Service</td>
</tr>
<tr>
<td>DTT</td>
<td>Digital Terrestrial Television</td>
</tr>
<tr>
<td>DVB-RCS/2</td>
<td>Digital Video Broadcasting – Return Channel via Satellite/2nd generation</td>
</tr>
<tr>
<td>DVB-S/2</td>
<td>Digital Video Broadcasting – Satellite/2nd generation</td>
</tr>
<tr>
<td>EMEA</td>
<td>Europe, Middle East and Africa</td>
</tr>
<tr>
<td>eMBB</td>
<td>Enhanced Mobile Broadband</td>
</tr>
<tr>
<td>ETSI</td>
<td>European Telecommunications Standards Institute</td>
</tr>
<tr>
<td>FR</td>
<td>Frequency Re-use factor</td>
</tr>
<tr>
<td>FSS</td>
<td>Fixed Satellite Service</td>
</tr>
<tr>
<td>GEO</td>
<td>Geostationary Earth Orbit</td>
</tr>
<tr>
<td>gNB</td>
<td>5G Node B (Base Station)</td>
</tr>
<tr>
<td>GSM</td>
<td>Global System for Mobile</td>
</tr>
<tr>
<td>GUI</td>
<td>Graphical User Interface</td>
</tr>
<tr>
<td>GW</td>
<td>Gateway (Satellite)</td>
</tr>
<tr>
<td>HTS</td>
<td>High Throughput Satellite</td>
</tr>
<tr>
<td>HTTP</td>
<td>HyperText Transfer Protocol</td>
</tr>
<tr>
<td>ICT</td>
<td>Information and Communications Technology</td>
</tr>
<tr>
<td>IP</td>
<td>Internet Protocol</td>
</tr>
<tr>
<td>ISDN</td>
<td>Integrated Services Digital Network</td>
</tr>
<tr>
<td>ISL</td>
<td>Inter-Satellite Link</td>
</tr>
<tr>
<td>KPI</td>
<td>Key Performance Indicator</td>
</tr>
<tr>
<td>L2S</td>
<td>Lower Layer Signalling</td>
</tr>
<tr>
<td>LEO</td>
<td>Low Earth Orbit</td>
</tr>
<tr>
<td>LNA</td>
<td>Low Noise Amplifier</td>
</tr>
<tr>
<td>LTE</td>
<td>Long Term Evolution</td>
</tr>
<tr>
<td>MABR</td>
<td>Multicast Adaptive Bit Rate</td>
</tr>
<tr>
<td>MANO</td>
<td>Management and Network Orchestration</td>
</tr>
<tr>
<td>MEC</td>
<td>Mobile Edge Computing</td>
</tr>
<tr>
<td>MEO</td>
<td>Medium Earth Orbit</td>
</tr>
<tr>
<td>MF-TDMA</td>
<td>Multi Frequency – Time Division Multiple Access</td>
</tr>
<tr>
<td>MIMO</td>
<td>Multiple-Input, Multiple-Output</td>
</tr>
<tr>
<td>mMTC</td>
<td>Massive Machine Type Communications</td>
</tr>
<tr>
<td>MPLS</td>
<td>Multi-Protocol Label Switching</td>
</tr>
<tr>
<td>MVE</td>
<td>Mobile Virtualised Equipment</td>
</tr>
<tr>
<td>Abbreviation</td>
<td>Full Form</td>
</tr>
<tr>
<td>--------------</td>
<td>-----------</td>
</tr>
<tr>
<td>TCP</td>
<td>Transmission Control Protocol</td>
</tr>
<tr>
<td>TG</td>
<td>Traffic Gateway</td>
</tr>
<tr>
<td>TN</td>
<td>Transport Network</td>
</tr>
<tr>
<td>TSSS or T3S</td>
<td>Traffic Steering, Splitting and Switching</td>
</tr>
<tr>
<td>TWTA</td>
<td>Traveling Wave Tube Amplifier</td>
</tr>
<tr>
<td>UC</td>
<td>User Control</td>
</tr>
<tr>
<td>UDM</td>
<td>Unified Data Management</td>
</tr>
<tr>
<td>UDR</td>
<td>User Data Repository</td>
</tr>
<tr>
<td>UE</td>
<td>User Equipment</td>
</tr>
<tr>
<td>UMTS</td>
<td>Universal Mobile Telephone System</td>
</tr>
<tr>
<td>UPC</td>
<td>User Plane Control</td>
</tr>
<tr>
<td>UPF</td>
<td>User Plane Function</td>
</tr>
<tr>
<td>VIM</td>
<td>Virtualised Infrastructure Manager</td>
</tr>
<tr>
<td>VM</td>
<td>Virtual Machine</td>
</tr>
<tr>
<td>VNF</td>
<td>Virtual Network Function</td>
</tr>
<tr>
<td>VPN</td>
<td>Virtual Private Network</td>
</tr>
<tr>
<td>VSAT</td>
<td>Very Small Aperture Terminal</td>
</tr>
<tr>
<td>WAN</td>
<td>Wide Area Network</td>
</tr>
</tbody>
</table>
1 Introduction

1.1 Document context

This deliverable D3.2, entitled “Integrated SaT5G Detailed Network Architecture” aims to provide advanced concept and design for integrated satellite-terrestrial 5G networks with consideration of the backhaul implementation options, the key technical enablers and the network management challenges. This first part of the deliverable (Part A) as the main output of WP3.2 – Backhaul Architectures, focuses on the detailed backhaul architecture based on the reference architecture defined in WP3.1 – Reference Satellite Network Architecture Integrated Into 5G (Figure 1-1).

![Diagram of WP3 Strategy approach and SWP3.x interaction](image)

This D3.2 part A therefore performs a review of existing techniques and technologies, mainly in the frame of satellite networks in order to identify different constraints and requirements for evolution and also disruptive solutions which will be taken into account for the design of the future integrated terrestrial-satellite network. Considering the backhaul implementation options defined in D3.1 which can be based on transport network or relay node approaches, and considering the on-going standardisation of 5G system, the following work is performed in this document, in line with the description of work:

- Identification of backhaul functionalities specific to satellite systems in integrated 5G architecture;
- Detailed definition of each identified functionality;
- Study of detailed backhaul architectures considering: Sat5G scenarios requirements (e.g. function delocalisation, management entities, security level, time to market, etc.), and the business models and identified technical criteria related to each implementation option;
- Analysis of potential impacts in 3GPP;
- Recommendation of the appropriate backhaul implementation architecture for a given Sat5G scenario;
Identification of relevant research topics to be further analysed in WP4 research pillars.

The topics analysed in the frame of this deliverable will be used as reference for potential contributions to standardisation organisations, and most importantly, for the development of technical solutions in the frame WP4.x, identified as key enablers for the satellite-terrestrial integration. Such developments can be related to requirements in term of virtualisation and orchestration which will be considered in WP4.1 and WP4.2, solutions for multilink management further addressed in WP4.3 and security aspects covered by WP4.4.

1.2 Organisation of the document

The document starts in chapter 2 with the state-of-the-art of satellite backhaul in order to identify the gaps and provide an overview of SaT5G reference architecture defined in D3.1 [2]. In this section, the reasons justifying the requirement for improvement of satellite backhauling techniques for the 5G are clearly exposed and the architecture considered as the baseline for the analysis is presented.

The analysis is therefore performed through the chapter 3 with the objectives of identifying the new requirements for satellite-terrestrial integration in term of high-level functionalities to be implemented either in the framework of satellite system, or in the framework of terrestrial network, considering the specificities of both networks and the on-going standardisation work.

In chapter 4, the backhaul implementation options based on relay nodes, as well as the implementation options based on the transport network are further specified, considering the high-level functionalities defined in chapter 3 in order to provide a strong basis for the analysis of the implementation option to be considered when performing the overall network design.

A role model defined in D2.3, the SaT5G use cases and scenarios defined in D2.1 and the backhaul implementation options defined in D3.1 are taken into account in chapter 4 to perform a network deployment analysis.

Chapter 5 provides the recommendations for standardisation work and for the WP4 research pillars and development work.

Finally, chapter 6 concludes the part A of the document.

The part B of the document deals with the caching and multicast architectures and its structure is presented in its dedicated introduction.
2 Generic backhauling architecture and implementation options

In terrestrial-mobile network, the backhaul network is the part of the network between the base station and the core network, in charge of routing the aggregated traffic from one end to another. The backhaul network is commonly based on terrestrial technologies such as optical fibre or point to point microwave. In the context of SaT5G, the considered backhaul network is the satellite backhaul based on telecommunication satellite systems used for various use case as investigated in D2.1 [3].

The objective of this section is to provide an overview of the state-of-the-art of satellite backhauling and identify the required limitations in order to specify the key solutions to enable efficient 5G satellite backhaul, compliant with most of the 5G requirements.

2.1 Satellite backhauling state-of-the-art

Mobile communication networks have evolved through three major generational leaps following technology trends and constantly evolving user demands. The first evolutionary jump was from the first generation, known as 1G, to the second generation, i.e. 2G, when the mobile voice network was digitised. The next evolutionary jump from 2G to 3G was made in order to fulfil the users’ ever-increasing demand for data and service quality. The proliferation of sophisticated user platforms, such as smart phones and tablets led to the next evolutionary leap towards 4G, which has made mobile networks provide a true wireless broadband service to its customers. With the enhanced options offered by 4G, new use cases, such as in health, automotive, entertainment, industrial, social, environmental etc. sectors, with diverse service requirements have been introduced. With such evolution and the Internet transforming towards an Internet-of-Things, mobile networks need to take their next evolutionary leap towards 5G, where a significantly greater volume of data will be delivered.

In major urban areas, the infrastructure to support all of this data is being built, but in rural areas, it is much harder for mobile network operators (MNOs) to make the case to provide backhaul to 4G and 5G base-stations. Some mobile spectrum licences have a coverage obligation attached, which increases the appetite for an economically workable rural backhaul solution. Using satellite links for such backhauling is an attractive option. However previous experience with satellite systems has led MNOs to perceive satellite backhauling services as islands of costly, difficult to integrate and inflexible connectivity. However, with the significant recent advances in satellite technologies that allow a lower cost per megabits and the advances in NFV and SDN technology which allow further flexibility of the satellite resources, satellites systems can overcome these limitations and MNOs can leverage satellite solutions more efficiently than before.

Satellite backhaul is an ideal choice for MNOs to expand their service in rural and remoted areas where the existing infrastructure is limited or non-existent. Deploying wired backhaul over difficult terrain like forests, mountains or deserts involves heavy investments and microwave backhaul requires meticulous planning and compliance with regulatory policies that take time and expense to put in place. In contrast, a satellite backhaul solution can be deployed with low cost of ownership, integrated with the terrestrial network to give good flexibility, and in a short time. The objective of reducing digital divide in area such as Europe would be therefore achievable with the concurrency of satellite system.

Even though satellite is naturally a well-suited solution for rural and remote areas, it still can bring benefits to urban areas in some cases. One example is a huge public gathering, where existing cellular network capacity may be temporarily overloaded, MNOs may choose to temporarily deploy base station sites to handle the additional data traffic. Another example is to provide resilience to the network if a base station has to be temporarily taken off air for maintenance purposes. Satellite backhaul enables rapid and easy commissioning of these temporary sites.

2.1.1 Backhaul via GEO satellites

Background:

GEO satellite has been used for many years to backhaul telecom services in some shape or form. They have evolved from providing analogue telephone circuits between continents in the 1970s using very large earth stations at both ends of the links, to now providing a wide range of digital services to
a wide range of satellite terminal types including VSATs and portable equipment. GEO satellites have also increased the range of spectrum they use; the earlier ones were usually C-band or Ku-band and the newer ones commonly utilise the Ka-band.

Today many operators market and provide cellular backhaul (2G, 3G, and now 4G) and Wi-Fi hotspot services as a quick check with your favourite search engine would show. This is supported by many of the satellite communications equipment vendors offering optimised configurations for their modems and satellite terminal equipment (often referred to as VSATs). Each different generation of cellular network presents its own set of problems that need to be carefully engineered. As the use cases and required bandwidths have increased (as illustrated below in Figure 2-1 [4]) so have the backhaul requirements.

<table>
<thead>
<tr>
<th></th>
<th>2G</th>
<th>2.5G</th>
<th>3G</th>
<th>4G</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Bandwidth</strong></td>
<td>64Kbps</td>
<td>2Mbps</td>
<td>200Mbps</td>
<td></td>
</tr>
</tbody>
</table>

**Figure 2-1: Evolving cellular use cases**

**Backhaul services today:**

Whilst some countries operators are demanding 3G and 4G satellite backhauls there are still parts of the world where they need satellite to extend their basic 2.xG cellular service reach. Providing cellular backhaul services by satellite is a major opportunity for the satellite communication industry.
The diagrams above (Figure 2-2 [5] and Figure 2-3 [6]) from Gilat and iDirect provide a useful insight into today's state of satellite backhaul connections. Other vendors and operators show similar diagrams. TCP throughput reduced with high round trip latency because of the slow growth in window size, and so GEO satellite links often employ a TCP acceleration algorithm that spoofs acknowledgement messages at each end and has a proprietary mechanism for error control.

Note that these diagrams both use traditional satellite terms such as VSAT and hub rather than the 3GPP-based terminology used elsewhere in the SaT5G project.

These and other vendor systems are used to provide cellular backhaul services – primarily to moving platform plus fixed rural and remote base stations. Use cases include permanent capacity, top-up capacity for rural cells in holiday areas, backup services and short term or temporary usage. One example of the latter is the services provided by Avanti to BT (now incorporating EE) for the UK Emergency Services Network (ESN, [7]) amongst other uses.

One of the main challenges in today's networks is that the traffic between the core and base station is encrypted. This means that, for example:

- To provide satellite TCP acceleration requires either that the operator is trusted, so it can break the VPN to provide acceleration (with the associated key management issues), or that satellite acceleration is provided with functions in the core and at the base station;
- CoS management across the satellite link can only be provided by the end-point devices unless either the satellite operator is trusted with the keys (see above) or CoS tags are provided on the packets outside the VPN tunnel (which is a non-standard operational mode).

A related sector is the provision of satellite connectivity to backhaul remote and moving Wi-Fi hotspots. This is less complicated to provide, as normal IP connections and techniques to manage end-user usage, access, and billing can be adapted from sectors such as in the hospitality sector.

### 2.1.2 Backhaul via non-GEO satellites

This section focuses on the MEO satellites and SES's O3b MEO system is taken as the example. In order to be able to provide coverage of a geographical region throughout the day, a constellation of MEO satellites needs to be in orbit so that at any given time at least one satellite is visible from that particular location, usually around 8 MEO satellites are needed and SES's O3b constellation has currently 16 in-orbit MEO satellites.

Figure 2-4 presents the high-level backhaul architecture when using a MEO satellite constellation. At the mobile site, a VSAT needs to be installed, enabling the communication with the satellite. The traffic from the base station will go through the satellite router to be modulated according to the air interface used, and transmitted via the remote antenna to the satellite. On the teleport site, after the traffic is received by the gateway antenna and demodulated from the satellite hub it is routed to the mobile core network through the satellite backbone fibre network.

It should be mentioned that during the backhauling via MEO satellite, both at the user site and the teleport, two antennas with tracking capability are needed; one that is communicating with the satellite which is currently serving the area and a second that is tracking the sky to connect with the next
satellite which will be serving the area. This handing over between satellites is one of the main differences to the GEO architecture.

![Figure 2-4: MEO based backhaul offerings](image)

Same paradigm is used for LEO constellations. These systems mainly differ from their operational orbit, LEO satellites are on a lower orbit than MEO satellites. But, they are still not fixed satellites and then imply the same issues than MEO satellites.

**Backhaul to the future:**

The use of satellite backhaul has sometimes been considered an expensive but necessary solution by MNOs. The lack of standardised Ethernet service support, including the visibility, control and performance usually associated with terrestrial backhaul solutions, inhibits an MNO’s ability to ensure consistent and uniform architectures and rollout planning. It also deprives operators of certainty in subscriber take-up rate, and a standard, high-quality experience across the entire network to lay the groundwork for increased Average Revenue Per User (ARPU).

By adopting industry standard Ethernet service constructs and orchestration, it is possible for a satellite-based backhaul solution to plug seamlessly into an MNO’s backhaul landscape – in the same manner as any terrestrial solution does. With the inter-carrier visibility and control of an automation solution based on Open Network Automation Platform (ONAP, coupled with the use of a lifecycle compliant with Metro Ethernet Forum (MEF) service orchestration (LSO), it is possible for an MNO to turn every user bit transported over satellite into a productive bit with no stranded capacity. Satellite players must ‘plug into’ the MNO ecosystem and become a true enabler of value-based outcomes.

As can be also seen from Figure 2-4, this the approach adopted by SES Networks. SES Networks is the only provider of managed data services over satellite in the world that can provide MEF Carrier Ethernet (CE 2.0) services. That means its Ethernet product has the same backhaul service attributes that MNOs purchase from terrestrial providers. MEF certification validates its delivery of fibre-equivalent connectivity.

### 2.1.3 Satellite backhaul access techniques

Star-topology VSATs, most of which using the DVB-S2X standard for the forward channel, are the most effective way to deliver satellite-based backhaul for cellular systems. The traditional deployment of satellite cellular is based on the Single Channel Per Carrier (SCPC) method.

SCPC means that a specific amount of satellite capacity is dedicated to one Base station, and that capacity is assigned to a single location whether it is in use or not – just like a leased line. If the network traffic for each Base station location is continuous, this type of setup works well because a limited amount of capacity is wasted. But if usage varies based on time of day, or the use of data services that are bursty by nature, then a dedicated link may be an expensive choice. The reason it is high-priced is that connectivity must be guaranteed during peak usage times, but outside of those times bandwidth that is paid for sits idle and is wasted. In areas where satellite bandwidth is in high demand, this can make using dedicated links an expensive and unacceptable proposition if trying to connect multiple Base station locations using SCPC.
A modern and cost-effective alternative to SCPC is to pool backhaul traffic onto a common network, replacing the previous standard SCPC links with a shared IP Time Division Multiple Access (TDMA) network. With this type of network, bandwidth is available to be shared dynamically among many different base station locations. TDMA allows operators to allocate satellite bandwidth according to the busy hour capacity requirements of the whole network rather than link-by-link. The end result is dramatically reduced bandwidth usage and lower costs.

In addition, using TDMA improves the flexibility and scalability of a network since the bandwidth pool is centrally managed. As the traffic from the remote Base station locations grows, the cellular operator only needs to increase the bandwidth pool to meet the overall demand of the expanding network. No technician is required to go to the remote site and change any settings. Instead, it is centrally managed. This allows an operator to start small then expand their network and increase bandwidth costs when subscribers are acquired and demand increases. Contrast this with SCPC where the operator must continuously re-balance or groom the capacity of all their links to try to keep up with changing patterns of usage.

Today, SCPC only makes economic or technical sense for very small networks; that is, those with only a handful of remote sites and relative low-speed links, where total transponder capacity requirements are low (e.g., just a few MHz) and opportunities for bandwidth sharing are limited. For larger networks (e.g., with >20 sites and many MHz of transponder capacity), the use of TDM/TDMA will result in much lower operating cost (OPEX) than incurred with SCPC technology. The relative financial advantage of TDM/TDMA vs. SCPC is shown in Figure 2-6.
In this cost analysis OPEX is the dominant consideration. The reason a TDM/TDMA network is not easily justified when there are just a few sites is due to the fixed capital cost of the Hub for a TDM/TDMA network, which is higher than the cost of a few SCPC modems. In the range of 20 to 50 sites, the total transponder OPEX, plus details of the network traffic patterns and various user requirements, must be examined more closely to determine which technology offers a lower cost of ownership.

For those sites that do require dedicated connectivity, almost all iDirect and Gilat modems can operate in both TDMA and SCPC modes. This means an operator can deploy a site using TDMA. If the site traffic grows to the point that it would justify dedicated capacity, it can change to SCPC mode on the same equipment without having to deploy a technician. The modem can be centrally controlled to switch over and operate as an SCPC modem to handle the dedicated traffic.

However, there are several improvements that can be done, which are studied in the SaT5G project. With the help of SDN, NFV and MANO the support of different services (slicing) can be obtained. The flexible bandwidth allocation to match network demand with available satellite capacity improves the network efficiency. Every bit transported over satellite is turned into a productive bit with no stranded capacity. Furthermore, MNOs can have a clear visibility on satellite-network resources, as well as the tools to leverage and manage them in a very flexible manner.

2.2 Requirements for 5G ready satellite systems

The satellite backhaul as presented in the section 2.1 above has always been a way to provide connectivity to isolated areas, but with limited capabilities in term of data rate or flexibility. As identified across different topics in section 2.1, an evolution of the satellite network will be necessary to support 5G requirements. The pre-analysis done in D3.1 [2] allowed to derive backhaul implementation options for the integrated satellite-terrestrial 5G network presented in section 2.4 below.

For previous satellite-terrestrial integration (earlier than 5G), the analysis have always being performed considerably after the specification of the targeted terrestrial system, yielding to a very limited integration between the networks. The 5G is an opportunity for definition of efficient integrated network but it comes with its own set of requirement that the satellite system will have to support.

In that sense, the analysis performed throughout this document tackles the following challenges:

- Introduce a standardised way to establish, manage operate non-terrestrial network either entirely by 3GPP NMS or in coordination with a satellite NMS;
- Introduce new solutions for satellite backhaul in 5G with the concept of RN for satellite backhaul and further specify the TN in standards;
- Support the requirement of flexible resource allocation, network slicing, mobility management;
- Adopt the SDN/NFV standardised interfaces and SatCom NFV deployment to allow efficient integration with terrestrial networks;
- Exploit multi-layer satellite systems (GEO, MEO, etc.) in order to be able to support various services and adjust the resources following the service requirements (e.g. latency, throughput, etc.).

The high-level features investigated in chapter 3 provide an analysis of the implementations required to face such challenges and the network deployment analysis in chapter 4 propose a way to implement such solution considering a role model and scenario requirements.

2.3 SaT5G reference architecture

The following diagram (Figure 2-7) shows the integrated terrestrial satellite architecture defined in the SaT5G deliverable D3.1 which is used as the basis for the analysis in this deliverable. The boxes with yellow background (SAT network functionalities, Hybrid multiplay functions, local DN) represent the main challenges addressed in our architecture design. The type of the satellite segment is not considered and it is assumed that the analysis performed throughout this document is applicable both to bent-pipe and to regenerative satellite system.
2.4 Satellite backhaul implementation options in 5G networks

2.4.1 General concepts

In integrated satellite-terrestrial network, the different backhaul (or indirect UE access) implementation options were detailed already in SaT5G deliverable D3.1 [2] and in ETSI documents [10]. Two major variants have emerged: transport network based implementation options, and relay node based implementation options. In each of the two variants, two sub-cases can be distinguished: either satellite access based on 3GPP defined protocols, or not. This approach leads to five different possible variants that are illustrated herein below (Figure 2-8). Relying on [2], the general concepts regarding transport network were already provided. For the sake of providing a self-contained output, the two options are illustrated and briefly discussed below in sections 2.4.2 and 2.4.3.
2.4.2 Relay node

A number of potential remote users could be located in areas with no connectivity to 5G services through any 5G terrestrial access networks, either temporarily due to the motion of the platform (aeronautical, maritime, or land, that is car or train) or due to economic circumstances.

In those cases, the purpose of using relays is to provide coverage extension at low cost to regions or locations where dedicated terrestrial backhaul links are not (or cannot be) deployed.

![Figure 2-9: SaT5G relay node approach](image)

The relay node is wirelessly connected (i.e. via satellite link or other wireless means) to the radio-access network via a donor cell.

Three relay node based implementation options for satellite backhauling have been identified in the framework of SaT5G, as described in [2]:

- Relay node based on Untrusted non-3GPP access;
- Relay node based on Trusted non-3GPP access;
- Relay node based on 3GPP access.

2.4.3 Transport network

As before, this backhaul implementation aims at providing 5G services to areas with no connectivity to 5G terrestrial networks.

The satellite gateway (GW) and UE are meant to establish the necessary communication link between the 5G mobile core network and the end-users’ UEs through the remote 5G RAN. By means of satellite terminology, the satellite UE is a modem that is deployed in the remote location, whereas the satellite gateway stands for the teleport that is part of the satellite ground infrastructure. Considering the case of a bent-pipe satellite, the space segment simply forwards the traffic to/from the forward/return link.

Two transport network based implementation options for satellite backhauling have been identified in the frame of SaT5G, as described in [2]:

- Transport network not based on 3GPP System Specifications;
- Transport network based on 3GPP System Specifications;

2.4.4 Analysis of the backhaul implementation options

The five identified satellite backhaul implementation options presented above throughout sections 2.4.2 and 2.4.3 are all candidates for 5G satellite backhauling with more or less the same level of support of 5G requirements. However, the implied requirements vary from one backhaul
implementation option to another, deriving to different level of integration, different types of interfaces and different time to markets, considering the maturity of the required technologies.

The Table 2-1 below, defined in D3.1 [2], recaps the key challenges for each backhaul implementation option and the time to market. This comparison is considered for the specification of the high-level features for 5G satellite terrestrial integration in chapter 3; it is also considered for the selection of candidate satellite backhaul implementation options for the network deployment analysis performed in chapter 4.

Table 2-1: Key challenges for the SaT5G backhaul implementation options and associated timescale

<table>
<thead>
<tr>
<th>Positioning</th>
<th>Backhaul implementation option</th>
<th>Key challenges</th>
<th>Network management</th>
<th>Potential additional supported features</th>
<th>Timescale</th>
</tr>
</thead>
<tbody>
<tr>
<td>Indirect Access (Backhaul)</td>
<td>Relay node with 3GPP Access</td>
<td>• NR over satellite</td>
<td>Single integrated NMS</td>
<td></td>
<td>Long Term</td>
</tr>
<tr>
<td></td>
<td>Relay node with Trusted non-3GPP Access</td>
<td>• Make satellite access a trusted non-3GPP access in standards</td>
<td></td>
<td>• Edge delivery</td>
<td>Mid Term</td>
</tr>
<tr>
<td></td>
<td>Relay node with Untrusted non-3GPP Access</td>
<td>• Implement untrusted access mechanisms as requested by 5G standards</td>
<td></td>
<td>• NF delocalisation</td>
<td>Short-Mid Term</td>
</tr>
<tr>
<td></td>
<td>Transport Network based on 3GPP System specification</td>
<td>• Design a specific “5G ready” satellite transport network based on 5G system specifications</td>
<td>3GPP NMS and Sat NMS working in coordination</td>
<td>• Hybrid multiplay (traffic steering at RAN level)</td>
<td>Short-Mid Term</td>
</tr>
<tr>
<td></td>
<td>Transport Network non-based on 3GPP System specification</td>
<td>• Design an adaptation layer for existing satellite transport network</td>
<td></td>
<td>• Enhanced UP, CP, MP interfaces between satellite domain and terrestrial domain</td>
<td>Short Term</td>
</tr>
</tbody>
</table>
3 High-level features applied to integrated satellite and terrestrial 5G network

SaT5G aims to define the detailed integrated satellite-terrestrial backhaul architectures and promote the satellite integration into 5G systems. As presented in Figure 2-7, the most important enablers for this seamless integration are the features implemented, mostly in the frame of satellite network, to allow the fulfilment of 5G requirements. This section looks forward to define these new high level features, which encompass various topics such as:

- Access management (e.g. registration, connection, roaming);
- End-to-End Quality of Service (e.g. adaptation between 5G standard and satellite capabilities);
- Mobility and handover management (e.g. beam and cell management, moving platform);
- Multilink implementation (e.g. AT3SF);
- Support of network slicing (e.g. coordination between NMS).

These suggested features are addressed in this chapter through the transport network (TN) and relay node (RN) implementation prism which aims to highlight the main architectural distinctions between these cases.

3.1 Registration, connection and roaming management

To provide connectivity to the UE, the access point (gNB or RN) needs to be itself registered and connected to the 5G core network as well as the satellite system. The connection of the access point can be therefore managed by the 3GPP NMS, typically for the RN based backhaul implementation options, reusing UE registration/connection procedure as defined in 3GPP TS 23.502 [11]. This aspect is further investigated in section 3.1.1.

For the transport network backhaul implementation options, the gNB need to be generic or transparent, in order to be able to connect to any MNO following a pre-established SLA (see section 3.1.2), especially for the moving platform backhaul use case where the access network will move across various areas covered by various SNO and MNOs.

The satellite terminal above the lower layers can be generic as well in order to be able to connect to any satellite system (e.g. GEO from different SNO, on different bands, MEO systems, etc.).

3.1.1 Registration and connection management

For backhaul implementation options based on relay node, as depicted in Figure 3-1 and evoked in section 5.1.4 of 3GPP TR 38874 [12], the RN can be registered and connected to the 5G core using UE registration and connection procedures defined in TS 23.502 [11]. The RN therefore acts as a proxy and forward the registration and connection requirements from the UE.

For the backhaul implementation options based on TN, it can be assumed that the SLA between the terrestrial network domain and the satellite domain includes that the satellite NMS have to share with the terrestrial 3GPP NMS the information about the status of the connectivity of satellite terminals, assuming that the gNB behind the satellite terminal is inherently identified and connected to the 5G system.

The satellite systems can use either legacy satellite registration with its connection procedures based on standards such as DVB S2/RCS2, or adopt the 3GPP NMS, taking advantage of the existing registration and connection procedures specified for terrestrial 3GPP systems.
3.1.2 Generic gNB-satellite terminal

Harmonisation of equipment in integrated satellite terrestrial network can be a key asset to provide flexibility at different levels and to make economy of scale.

As depicted in the Figure 3-2, the analysis of this harmonisation is performed from 2 perspectives: the generic satellite terminal and the generic gNB.
**Generic satellite terminal:**

The satellite terminal (represented as NTN terminal in the Figure 3-2) needs to be generic and flexible in order to be able to connect to various satellite systems, including various orbits and technologies, from different SNOs. The adoption of SDN/NFV paradigms and the development of Software Defined Radio (SDR) based equipment will be necessary to enable such flexibility; as will electrically or electronically steerable antennas. It is likely that achieving such flexibility will take years of development; the first step can be the flexibility in term of ability to connect to any GEO satellite from any SNO.

**Generic gNB:**

We can introduce in 5G systems the concept of generic gNB that can be connected to different MNOs. In this concept the gNB does not belong to a specific MNO but should be able to connect to any.

This is a kind of roaming scenario, but at RAN level rather than the UE level; and without a HPLMN (since the gNB / RN does not belong to a specific MNO).

This kind of generic equipment can typically be embedded on moving platform with gNB functionalities providing 5G connectivity to the passengers. This equipment could be operated by the moving platform operator, with a selection of SNO which cover the moving platform route, and a selection of MNO’s to provide a 5G core. The associated negotiation process can be performed by a third party (concept of broker further elaborated in section 4.1).

### 3.2 Resource allocation and orchestration

Resource allocation is a key issue in 5G, as traffic growth is rapid. Problems of resource allocation occur at a wide range of scales (both time scales and traffic scales), and affect resource blocks in the air interface, and spectrum purchase and brokerage.

To address this topic, the concept of a broker discussed in WP2 business modelling is evoked hereafter and as the basis of Sat5G approach to handle resource allocation and orchestration. In this context, the Satellite Network Operator (SNO) uses the satellite system infrastructure and markets its capacity to MNOs or third parties such as brokers.

As discussed in SaT5G deliverable D2.3 [13], it is proposed that each SNO runs a component called a Resource Manager function (RM) which keeps track of all available resources, collects usage, and makes decisions. The RM can be either centralised or decentralised. If the management functions and decision-making used to allocate resources are concentrated at a central component, there can be problems of lack of scalability when the system size increases. The centralised RM can get heavily loaded and become a bottleneck. The idea of Decentralised Resource Manager (dRM) is that individual management units are able to acquire information about the state of the entire system by communicating their local information to all others. As a result of decentralised management the decision-making process is dispersed among different management components so as to alleviate the problems encountered in a centralised approach.

When a base station requires a network service for its customer, the RM will choose a good backhaul resource for this service. If the choice is a non-terrestrial resource, the MNO will do a request through a standard template with the requirements of the new service in term of bandwidth, availability, QoS, area(s) to provide the service, mobility, busy hour, maximum amount of dynamic resources, etc. This template is sent to a broker. From a business point of a view, a MNO may not be interested in managing a satellite network. For that reason, a broker is a relevant figure that provides satellite capabilities to MNOs and, at the same time, ensures the satellite utilisation to SNOs. A broker provides the best available dynamic satellite resources to the MNO. This typically works for the backhaul implementation options based on satellite as transport network.

Several MNOs can create a group to trade resources together with SNOs when they have common needs. This new association does not preclude individuals or subgroup trades. Their demand will then pass through an element called Joint Resource Management (JRM). The JRM works in the same way as the RM. Once the JRM receives the response from the broker to its requests, it forwards them to the RMs subscribed to this trade and then individually analysed by each MNO.
Just as an MNO, SNOs have their own RM and for security and privacy, an abstraction layer provides isolation of their core network from the third parties. This abstraction layer communicates with its RM, which requests and gets data under its broker request. Through its RM, the SNO and the broker starts a process to exchange information until the broker has all the requested data. In parallel, the broker runs the same process with different SNOs and brokers.

3.3 Quality of service

High speed and high throughput, high reliability, low latency, high capacity, high availability, high connectivity and dynamic bandwidth allocation are the key QoS benefits that the 5G systems expect. First, this section aims to introduce the legacy of satellite Class of Service (CoS) and an overview of the QoS treatment over 5G network. Then, 5G QoS model adaption over satellite will be addressed with respect to the backhaul implementation options based on RN and TN.

3.3.1 Legacy Sat QoS and requirement for evolution

DiffServ approach:

The propagation delay in the satellite link being higher compared with terrestrial links, the delays can contribute to a major problem, for example, in an acknowledgment and time-out based congestion control mechanism (like TCP), where the performance is inherently related to the delay-bandwidth product of the connection. TCP round-trip time (RTT) measurements are sensitive to delay variations which may cause false timeouts and retransmissions. As a result, the congestion control issues for broadband satellite networks are somewhat different from those of low-latency terrestrial networks.

One mechanism to meet the demand for IP QoS over satellite networks is the Differentiated Services (DS) architecture. DS provides scalable service differentiation in the Internet that is used to permit differentiated pricing of Internet service. This differentiation may either be quantitative or relative. DS is scalable as traffic classification and conditioning is performed only at network boundary nodes. The service to be received by traffic is marked as a code point in the DS field in the IPv4 or IPv6 header. The DS code point in the header of an IP packet is used to determine the Per-Hop Behaviour (PHB), i.e. the forwarding treatment it will receive at a network node. Currently, formal specification is available for two PHBs - Assured Forwarding and Expedited Forwarding. In Expedited Forwarding, a transit node uses policing and shaping mechanisms to ensure that the maximum arrival rate of traffic aggregate is less than its minimum departure rate. At each transit node, the minimum departure rate of traffic aggregate should be configurable and independent of other traffic at the node. Such per-hop behaviour results in minimum delay and jitter and can be used to provide an end-to-end `Virtual Leased Line' type of service.

In Assured Forwarding (AF), IP packets are classified as belonging to one of four traffic classes. IP packets assigned to different traffic classes are forwarded independently of each other. Each traffic class is assigned to a minimum configurable amount of resources (link bandwidth and buffer space). Resources not being currently used by another PHB or an AF traffic class can optionally be used by remaining classes. Within a traffic class, a packet is assigned to one of three levels of drop precedence (green, yellow, red). In case of congestion, an AF-compliant DS node drops low precedence (red) packets in preference to higher precedence (green, yellow) packets.

QoS management in a state-of-the-art VSAT network:

Modern VSAT (Very small aperture terminal – the common vernacular for satellite terminal) networks support extensive methods for managing QoS. The determination of QoS can be made using one or more of the following criteria:

- Source or destination IP address;
- Source or destination TCP or UDP socket;
- IP CoS (DSCP Flag - differentiated services code point - often referred to as DiffServ);
- VLAN tag;
- A few high end systems also support a limited deep packet inspection capability to determine protocol type.

Some VSAT systems can also support L2 traffic and if allowed can determine these criteria within the L2 packets. VSAT networks often support multiple virtual networks; each within their own bandwidth pool. These tools require the VSAT hub and terminal having sufficient visibility to be able to identify
the packet and send it to the required queue. If the data is encrypted within an IPsec tunnel then either; (a) a copy of the keys is provided to allow the system visibility to identify the packet priority (i.e. sufficient trust is provided to have the right keys for this purpose; (b) the DCSP flag is set correctly; or (c) the data is sent at the default priority. Packets sent using either unencrypted tunnelling (such as GRE) or encryption (e.g. TLS) can be better identified.

These criteria can be used to define, depending on VSAT type and manufacturer, a series of different service classes implemented using the following tools:

- Application DiffServ QoS across single and/or multiple remotes sites;
- Class-based Queuing – Assign percentage of bandwidth to each class, bandwidth pools for defined classes within defined groups of VSATs;
- Adaptive Constant Bit Rate with Minimum, Maximum, and user-definable step-sizes;
- Backlog-based dynamic stream with weighted fair queuing;
- Class-based weighted prioritisation;
- Rate Limiting – Allocate only bandwidth that is needed;
- Committed Information Rate – Dedicate bandwidth as required (Static and Dynamic) with Minimum, Guaranteed, and Maximum rates.

Some VSATs can switch return channel mode to switch between optimum support for normal bursty traffic, and to support high bandwidth and/or low jitter applications.

The rules are defined on the VSAT network management system and implemented in real time by a) a network function such as a network control centre in the VSAT hub for traffic in the forward link, and b) by the VSAT in coordination with the network control centre.

Typically these VSAT systems support four priority levels of IP based traffic, with a fifth higher level offering constant bit-rate. Some vendors implement IP multicast traffic in one of the IP traffic levels, others create an additional level for multicast only.

The following Table 3-1 indicates where a physical link KPI has impact on service class.

<table>
<thead>
<tr>
<th>Service Classes</th>
<th>Latency</th>
<th>Packet Error Rate</th>
<th>Data Rate</th>
<th>Priority level</th>
<th>Jitter</th>
<th>Constant bit rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Application DiffServ QoS</td>
<td>Y</td>
<td>(1)</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>Class Based Queuing</td>
<td>Y</td>
<td>(1)</td>
<td>y</td>
<td>y</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>Adaptive Constant Bit Rate</td>
<td>Y</td>
<td>(1)</td>
<td>y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>Backlog-based dynamic stream</td>
<td>Y</td>
<td>(1)</td>
<td>Y</td>
<td>Y</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Class-based weighted prioritisation</td>
<td>Y</td>
<td>(1)</td>
<td>Y</td>
<td>Y</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rate Limiting</td>
<td>(1)</td>
<td></td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td></td>
</tr>
<tr>
<td>Committed Information Rate</td>
<td>(1)</td>
<td></td>
<td>Y</td>
<td>Y</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Notes:

(1) Satellite links generally operate in quasi-error-free mode for unicast traffic, therefore the service class selection can have little relevance.

(2) Latency on a correctly dimensioned satellite system tends to be somewhat fixed (the satellite path length dominating, with queuing delays secondary), however higher priority packets will have smaller queuing delays unless that higher priority bandwidth pool is overloaded.

Requirements for evolution:
The following features are seen as beneficial for 5G terrestrial-satellite interworking:

- Standards based key management to allow access to IPsec keys where sufficient trust exists;
- Clear standards based mapping between satellite CoS and 5G classes (5G QI);
- Process to accept slice request with 5G QI enabling a feedback mechanisms (e.g. to state the requested criterion is not met).

3.3.2 Overview of QoS in 5G network overview

In 5G, the QoS Flow is the finest granularity of QoS differentiation in the PDU Session. A QoS Flow ID (QFI) is used to identify a QoS Flow in the 5G System. User-Plane traffic with the same QFI within a PDU Session receives the same traffic forwarding treatment. The QFI is unique within a PDU session and is carried as an encapsulation header on N3 (and N9) without any changes to the E2E packet header.

Any QoS Flow is characterised by (TS 23.501 [5.7.1]):

- A QoS profile provided by the SMF to the AN via the AMF over the N2 reference point or preconfigured in the AN;
- One or more QoS rule(s) which can be provided by the SMF to the UE via the AMF over the N1 reference point and/or derived by the UE by applying reflective QoS control\(^1\);
- One or more SDF template(s) provided by the SMF to the UPF.

---

\(^1\) Reflective QoS control: UE applies the same QoS received without any further exchange.
Figure 3-3 depicts the structure of the 5G QoS model where the main object is the QoS Flow which is entirely defined by a set of parameters composed by a QoS Profile, a QoS Rule and a SDF template. For further details on the structure of the 5G QoS model, refer to TS 23.501 [14]. This 5G QoS model is used to classify the user-plane traffic and to map the QoS flows to the access network resources.

Figure 3-4: Classification, marking and mapping of the user plane traffic to access network resources

Figure 3-4 [14] depicts the positioning of each 5G QoS model elements. For the DL, data packets are classified by UPF using SDF templates to generate QoS Flows. These QoS Flows are then mapped to AN resources. For the UL, data packets are mapped to QoS Flows and those are marked using QoS Rules. These QoS Flows are then mapped to AN resources.

Table 3-2: 5G QoS model acronyms

<table>
<thead>
<tr>
<th>Information</th>
<th>Description</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>QoS characteristics</td>
<td>Specific parameters (resource type, priority, delay, packet error rate, etc.) for each QoS flow</td>
<td>-</td>
</tr>
<tr>
<td>5QI</td>
<td>A 5G QoS parameter, combination of 5G characteristics</td>
<td>-</td>
</tr>
<tr>
<td>5G QoS flow</td>
<td>Finest granularity of QoS differentiation in the PDU Session and characterised by PDR(s) and QoS rule(s)</td>
<td>-</td>
</tr>
<tr>
<td>QFI</td>
<td>Identify the QoS flow</td>
<td>SMF</td>
</tr>
<tr>
<td>5G QoS profile</td>
<td>Performance expectations placed on the 5G QoS flow. Based on QoS parameters</td>
<td>SMF</td>
</tr>
<tr>
<td>5G QoS parameters</td>
<td>Parameters like 5QI, ARP, ROA that produce the QoS profile</td>
<td>-</td>
</tr>
<tr>
<td>5G QoS rules</td>
<td>Classification and marking of UL User plane traffic in order to associate it with a QoS flow</td>
<td>SMF</td>
</tr>
<tr>
<td>Packet Filter Set</td>
<td>Used in the QoS rules to identify one or more packet (IP or Ethernet) flows</td>
<td>-</td>
</tr>
<tr>
<td>Packet Detection Rule</td>
<td>Contains information to classify a packet arriving at the UPF</td>
<td>SMF</td>
</tr>
</tbody>
</table>

The main concept of the 5G QoS model is the 5G QoS Flow since it is the finest QoS granularity level. However, it is not directly related to the physical parameters which are gathered under the QoS characteristics terminology. This role is endorsed by the 5QI which is one of the QoS Profile settings.

The 5G QoS characteristics (section 5.7.3 of TS 23.501 [2]) describe the packet forwarding treatment that a QoS Flow receives edge-to-edge between the UE and the UPF in terms of the following performance characteristics:

- Resource Type: indicates QoS Flow level related to network resources, which can be Guaranteed Bit Rate (GBR), GBR Delay Critical and Non-GBR;
- Priority level: indicates a priority in scheduling resources among QoS Flow Priority. It is signalled with standardised 5QIs, and if it is received, it overwrites the default value specified in QoS characteristics. The lowest Priority level value corresponds to the highest Priority;
- Packet Delay Budget (PDB): defines an upper bound for the time that a packet may be delayed between the UE and the UPF. For a delay-critical GBR resource type, packets delayed more than the PDB are added to the PER;
- Packet Error Rate (PER): defines an upper bound for a rate of non-congestion-related packet losses that have been processed by the sender of a link layer protocol (e.g. RLC in RAN of a 3GPP access) but that are not successfully delivered by the corresponding receiver to the upper layer (e.g. PDCP in RAN of a 3GPP access);
- Averaging window (defined only for GBR QoS Flows): represents the duration over which the GFBR and MFBR shall be calculated;
- Maximum Data Burst Volume (for 5QIs with 5G Access Network PDB <=20ms): denotes the largest amount of data that the 5G-AN is required to serve within a period of 5G-AN PDB (i.e. 5G-AN part of the PDB).

The QoS characteristics are considered in 3GPP as guidelines for setting node-specific parameters for each QoS Flow.

A combination of QoS characteristics leads to a 5QI. Standardised 5QI values have one-to-one mapping to a standardised combination of 5G QoS characteristics as specified on Table 3-3.

<table>
<thead>
<tr>
<th>5QI</th>
<th>Resource Type</th>
<th>Default Priority Level</th>
<th>Packet Delay Budget (ms)</th>
<th>Packet Error Rate</th>
<th>Default Maximum Data Burst (B) Volume</th>
<th>Default Averaging Windows (ms)</th>
<th>Example services</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>GBR</td>
<td>20</td>
<td>100</td>
<td>0.01</td>
<td>N/A</td>
<td>2000</td>
<td>Conversational voice</td>
</tr>
<tr>
<td>2</td>
<td>GBR</td>
<td>40</td>
<td>150</td>
<td>0.01</td>
<td>N/A</td>
<td>2000</td>
<td>Conversational video (live streaming)</td>
</tr>
<tr>
<td>3</td>
<td>GBR</td>
<td>30</td>
<td>60</td>
<td>0.01</td>
<td>N/A</td>
<td>2000</td>
<td>Real time gaming, V2X messages, electricity distribution, medium voltage, process automation, monitoring</td>
</tr>
<tr>
<td>4</td>
<td>GBR</td>
<td>50</td>
<td>300</td>
<td>0.000001</td>
<td>N/A</td>
<td>2000</td>
<td>Non-conventional video (buffered streaming)</td>
</tr>
<tr>
<td>55</td>
<td>GBR</td>
<td>7</td>
<td>75</td>
<td>0.01</td>
<td>N/A</td>
<td>2000</td>
<td>Mission critical user plane Push To Talk voice (MCPTT)</td>
</tr>
<tr>
<td>66</td>
<td>GBR</td>
<td>20</td>
<td>100</td>
<td>0.01</td>
<td>N/A</td>
<td>2000</td>
<td>Non-mission critical user plane Push To Talk voice (MCPTT)</td>
</tr>
<tr>
<td>67</td>
<td>GBR</td>
<td>10</td>
<td>100</td>
<td>0.01</td>
<td>N/A</td>
<td>2000</td>
<td>Mission critical video user plane</td>
</tr>
<tr>
<td>75</td>
<td>GBR</td>
<td>25</td>
<td>50</td>
<td>0.01</td>
<td>N/A</td>
<td>2000</td>
<td>V2X messages</td>
</tr>
<tr>
<td>5</td>
<td>non-GBR</td>
<td>10</td>
<td>160</td>
<td>0.000001</td>
<td>N/A</td>
<td>N/A</td>
<td>IMS signalling</td>
</tr>
<tr>
<td>6</td>
<td>non-GBR</td>
<td>60</td>
<td>300</td>
<td>0.000001</td>
<td>N/A</td>
<td>N/A</td>
<td>Video (Buffered Streaming) TCP-based (e.g., www, e-mail, chat, ftp, p2p file sharing, progressive video, etc.)</td>
</tr>
<tr>
<td>7</td>
<td>non-GBR</td>
<td>70</td>
<td>100</td>
<td>0.01</td>
<td>N/A</td>
<td>N/A</td>
<td>Voice, Video (Live Streaming) Interactive Gaming</td>
</tr>
<tr>
<td>8</td>
<td>non-GBR</td>
<td>80</td>
<td>300</td>
<td>0.000001</td>
<td>N/A</td>
<td>N/A</td>
<td>Video (Buffered Streaming) TCP-based (e.g., www, e-mail, chat, ftp, p2p file)</td>
</tr>
<tr>
<td>9</td>
<td>non-GBR</td>
<td>90</td>
<td>60</td>
<td>0.000001</td>
<td>N/A</td>
<td>N/A</td>
<td>Video (Buffered Streaming) TCP-based (e.g., www, e-mail, chat, ftp, p2p file)</td>
</tr>
<tr>
<td>69</td>
<td>non-GBR</td>
<td>5</td>
<td>200</td>
<td>0.000001</td>
<td>N/A</td>
<td>N/A</td>
<td>Mission Critical delay sensitive signalling (e.g., MC-PTT signalling)</td>
</tr>
<tr>
<td>70</td>
<td>non-GBR</td>
<td>55</td>
<td>50</td>
<td>0.000001</td>
<td>N/A</td>
<td>N/A</td>
<td>Mission Critical Data (e.g., example services are the same as QCI 5/8/9)</td>
</tr>
<tr>
<td>79</td>
<td>non-GBR</td>
<td>65</td>
<td>10</td>
<td>0.01</td>
<td>N/A</td>
<td>N/A</td>
<td>V2X messages</td>
</tr>
<tr>
<td>80</td>
<td>Delay Critical</td>
<td>68</td>
<td>0.000001</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>Low Latency eMBB applications Augmented Reality</td>
</tr>
<tr>
<td>81</td>
<td>GBR</td>
<td>11</td>
<td>5</td>
<td>0.00001</td>
<td>160</td>
<td>2000</td>
<td>Remote control (see TS 22.261)</td>
</tr>
<tr>
<td>82</td>
<td>GBR</td>
<td>12</td>
<td>10</td>
<td>0.00001</td>
<td>320</td>
<td>2000</td>
<td>Intelligent transport systems</td>
</tr>
<tr>
<td>83</td>
<td>GBR</td>
<td>13</td>
<td>20</td>
<td>0.00001</td>
<td>640</td>
<td>2000</td>
<td>Intelligent transport systems</td>
</tr>
<tr>
<td>84</td>
<td>GBR</td>
<td>19</td>
<td>10</td>
<td>0.00001</td>
<td>256</td>
<td>2000</td>
<td>Discrete automation</td>
</tr>
<tr>
<td>85</td>
<td>GBR</td>
<td>22</td>
<td>10</td>
<td>0.0001</td>
<td>1358</td>
<td>2000</td>
<td>Discrete automation</td>
</tr>
</tbody>
</table>

Some details on the QoS management are provided in section 3.3.3 with an analysis of their application to the backhaul via relay node implementation options. Complete specifications are provided in 3GPP TS 23.501 [14] and [15].
3.3.3 QoS management for relay node implementation options

The management of QoS in 5G can typically be adopted or adapted for the RN. However, this management has to take into account the satellite system specifications. This paragraph aims to describe a QoS adaptation mechanism for RN based backhaul implementation options while 3.3.4 focuses on TN based backhaul implementation options.

The basic architecture is described in TS 38.300 and can be applied to NR and E-UTRA connected to the 5GC. First, 5GC establishes a PDU session for each UE and the NG-RAN establishes at least one Data Radio Bearer (DRB) per PDU session. The NG-RAN can also configure several DRBs for others QoS flows. Then, the NG-RAN maps the packets from the right PDU session to the right DRB with two logical steps:

- At Non-Access Stratum (NAS) level, the packet filters (UE and 5GC) associates packets to the right QoS flows
- At Access Stratum (AS) level, the mapping rules (UE and 5GC) associates QoS flows to the right DRBs

These two steps ensure quality of service from IP packets to a DRB. At NAS level, all QoS flow characteristics are defined in the 3.3.2. At AS level, the mapping of QoS flows to DRBs is managed by the mapping rules. These mapping rules can be signalled by two ways:

- Reflective mapping: UE monitors the QFI (provided by UPF to RAN) of the downlink packets, for each DRB, and applies the same mapping for the uplink
- Explicit configuration: the mapping rules are signalled by RRC

SDAP (Service Data Adaptation Protocol) is responsible for mapping QoS flows on DRBs, marking the QFI in both DL and UL packets and mapping reflective QoS flows to DRBs for the UL SDAP PDUs (TS 37.324). This protocol is actually the first step of an adaptation function but this approach is static.

In this context, from a SaT5G framework perspective, a dynamic QoS adaptation is foreseen to mitigate the satellite link potential variations (e.g. propagation delay, fading, etc.). This dynamic adaptation should also improve flexibility and agility in traffic management (e.g. prioritisation, load balancing, etc.) within the system. One potential solution can be provided by RRC since SDAP can be dynamically reconfigured by the RRC protocol.

Figure 3-6 depicts the positioning of the SDAP protocol. The SDAP protocol could be reconfigured dynamically by RRC signalling on both sides (NTN GW and NTN terminal). This mechanism would allow the management system to update the mapping table between the 5QI and the satellite resources. Therefore, the solution proposed consists of an updating of this mapping table by the control plane via RRC signalling to SDAP protocol.
An example of a mapping table is given on Table 3-4. This table needs to be further investigated in the frame of SDOs in order to identify the relevant parameters to be specified and associated range of values.

<table>
<thead>
<tr>
<th>5QI</th>
<th>Satellite Resources</th>
</tr>
</thead>
<tbody>
<tr>
<td>non-GBR</td>
<td>{5, 6, 8, 9, 70} High SNR, robust Modcod, high bandwidth</td>
</tr>
<tr>
<td></td>
<td>7 low SNR, high bandwidth, high Modcod</td>
</tr>
<tr>
<td></td>
<td>79 low SNR, low bandwidth, low Modcod</td>
</tr>
<tr>
<td>{69, 80}</td>
<td>TBD</td>
</tr>
<tr>
<td>GBR</td>
<td>{1, 65, 66, 75} TBD</td>
</tr>
<tr>
<td>{2, 3, 67}</td>
<td>TBD</td>
</tr>
<tr>
<td>4</td>
<td>TBD</td>
</tr>
<tr>
<td>TBD</td>
<td>{81, 82, 83, 84, 85} TBD</td>
</tr>
</tbody>
</table>

### 3.3.4 QoS management for Transport Network implementation options

This section will focus on the QoS adaptation mechanism for TN implementation options. All the concepts related to the QoS for RN implementation options are applicable here. Figure 3-7 depicts the configuration which can be encountered by the QoS management in the case of TN implementation options. From a schematic point of view, there are two entities needed to achieve a QoS adaptation between the 5GCN and the satellite system: sat CoS/Sat resources and 5QI/sat CoS.
Figure 3-7: QoS adaptation configuration for TN implementation options

The adaptation between 5QI and CoS is the key issue for integrating the satellite into 5G networks, because the mapping between the sat CoS and the satellite resources is already done in current SatCom systems (i.e. SNO maps CoS on its satellite resources).

Figure 3-8: SDAP positioning illustration

As shown in Figure 3-8, on solution for QoS adaptation in TN implementation options can be the adoption of SDAP. The adaptation mechanism is similar to the RN approach presented in section 3.3.3. The RRC signalling can therefore configure dynamically the SDAP protocol at the NTN GW and NTN terminal. This implies the support of these protocols at NTN GW and NTN terminal level. This is the same positioning that the one for RN implementation options, but there are two adaptation layers here. The mapping between 5QI and satellite CoS is the most challenging one because of the granularity gap. There is no bijection between 5QI and sat CoS (the same issue appears when dealing with DCSP code point and IP precedence mapping table). At present, there are 22 5QI standardised but only 8 level of CoS.
Table 3-5: Example of mapping table between 5QI and sat CoS

<table>
<thead>
<tr>
<th>Sat CoS</th>
<th>5QI</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>{5, 6, 8, 9, 70}</td>
</tr>
<tr>
<td>1</td>
<td>non-GBR</td>
</tr>
<tr>
<td>2</td>
<td>7</td>
</tr>
<tr>
<td>3</td>
<td>79</td>
</tr>
<tr>
<td>4</td>
<td>{69, 80}</td>
</tr>
<tr>
<td>5</td>
<td>GBR</td>
</tr>
<tr>
<td>6</td>
<td>4</td>
</tr>
<tr>
<td>7</td>
<td>TBD</td>
</tr>
</tbody>
</table>

Table 3-6: Example of mapping table between Sat CoS and satellite resources

<table>
<thead>
<tr>
<th>Sat CoS</th>
<th>Satellite Resources</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>High SNR, robust Modcod, high bandwidth</td>
</tr>
<tr>
<td>1</td>
<td>low SNR, high bandwidth, high Modcod</td>
</tr>
<tr>
<td>2</td>
<td>low SNR, low bandwidth, low Modcod</td>
</tr>
<tr>
<td>3</td>
<td>TBD</td>
</tr>
<tr>
<td>4</td>
<td>TBD</td>
</tr>
<tr>
<td>5</td>
<td>TBD</td>
</tr>
<tr>
<td>6</td>
<td>TBD</td>
</tr>
<tr>
<td>7</td>
<td>TBD</td>
</tr>
</tbody>
</table>

Table 3-5 and Table 3-6 illustrate what could be the mapping tables between Sat CoS and 5QI and between Sat CoS and satellite resources. One solution consists of an updating of these mapping tables by the control plan via RRC signalling to SDAP protocol. This solution has to be further studied in terms of RRC and SDAP protocols and related procedures.

3.4 (Radio) Access Network aspects

3.4.1 Mobility management in 3GPP Release 15

Mobility management is tightly related to the connectivity status of a UE to the serving base station. Typically, the Radio Resource Control (RRC) configuration modes provide information about the status of a UE when it is connected to the network. Besides typical UE status like RRC Connected and RRC Idle modes, in Release 15, 3GPP has introduced a new RRC state called RRC Inactive, which is necessary to support not only eMBB services but also uRLLC and mMTC. From the perspective of mobility, the session management for UEs in active mode (i.e., RRC Connected) is the most important since it requires relocating the current data bearer(s) to a target base station in a seamless manner, while RRC Idle mode UEs just go through a cell reselection process [15]. Therefore, mobility management will be discussed in this section for a UE that holds an active data session.

Mobility management can be distinguished also based on the availability of Xn interface between base stations (i.e. gNB) or not. In case the Xn interface is available, there will be an Xn based handover, whereas in the opposite case an N2 handover will be executed. Independently on the availability of the Xn interface, the typical handover procedure includes the following phases:

- UE measurements report to the serving gNB;
- Handover preparation;
- Handover execution (UE connects to the target gNB);
- Handover completion (data path switch at the 5G CN and UE context release).

The 3GPP Release 15 defined Xn based handover procedure for a UE in 5G is shown in Figure 3-9. A UE constantly measures signals from neighbouring base stations, doing intra- and inter-frequency measurements on dedicated carriers. When a mobile terminal moves away from the serving cell, it will receive progressively lower power until falls below the power received from a neighbouring cell.
(hysteresis threshold). In case this condition persists for at least a period denoted by Time-To-Trigger (TTT), a handover procedure is commenced. Measurements are reported to the serving gNB (referred to also as Source gNB) that can decide whether to begin a handover or not selecting the best target base station (i.e. Target gNB) amongst the possible neighbours relying on the measurement reports of the UE. This is done with the serving gNB that sends a request directly to the Target gNB. The selected cell performs an admission control procedure to decide whether to accept or reject a new UE. In case the UE is rejected (e.g. if the target cell is overloaded), another target cell has to be selected, which might not be the best possible in terms of radio signal reception. When the target cell can accept the new UE, the target gNB will reply to the request with an acknowledgement and the handover begins with typical RRC reconfiguration request sent to the UE. As soon the handover preparation phase ends, the handover execution phase starts. During execution, the UE synchronises with the target cell and the user plane traffic is transferred to the target gNB. The handover is terminated by the handover completion phase in which the data path switch is completed over the N2 interface, the UE context transferred to the target gNB and resources released from the origin base station. It is worth noticing the interaction between SMF and UPF over the N4 interface for modifying the PDU session.

**Handover undesired effects:**

It must be noted that the handover process (either Xn based or N2 based) is error-prone, and several issues may still arise. The most well-known problems include:

- **Handover failure:** in this case the UE does not manage to complete the handover successfully and the connection is dropped before this is attached to the new cell. This leads to a condition of radio link failure and consequently the start of a new cell reselection procedure, which implies connection drop that must be followed by a new attach phase carried out by the UE.

- Among the possible events that cause handover failure, it is relevant to mention:
  - **Too early handover:** this implies a situation of radio link failure at the target cell since the radio signal of the target base station is too low or not constant to maintain the UE session alive.
  - **Too late handover:** this implies a situation of radio link failure at the serving cell since the UE loses the connection to the origin base station prior to completing the handover procedure towards the target cell.

- For completeness, it is also worth mentioning the ‘Ping-Pong’ effect, a type of problem that arises when a UE is handed over between two adjacent cells causing additional signalling to be exchanged and likelihood to incur in a radio link failure.

The problems illustrated so far are common to all generations of mobile network technology and are not specific to 5G. Possible well-known remediation in the existing literature includes optimisation of the handover, especially of the handover hysteresis and TTT parameters. Moreover, self-organising features such as Mobility Robustness Optimisation (MRO) techniques have been the subject of intense research for several years in 4G. Providing adaptation of the handover parameters can lead to either handover anticipation or postponement depending on the specific strategy to adopt. In other words, the cell size becomes the objective of optimisations to minimise the connection drops.

For the sake of completeness, it is worth noticing that additional challenges are introduced to mobility management in the case that source and target cells use different backhaul technologies. Specifically, if the handover takes place to a target cell that relies on a satellite backhaul optimisation of the procedure is required to relieve connection drops that could arise due to the satellite link delay. For instance, referring to Figure 3-9, the phase in which data is forwarded and buffered at the target gNB will be affected by the potentially long latency of the backhaul link. In this regard, handover anticipation or even early data forwarding between source and target cells can be envisaged. In the latter case, tight coordination between cells is likely to be required borrowing even from other 3GPP features such as coordinated multipoint transmissions. In overall, new MRO techniques that are designed to optimize the handover procedure in the presence of long backhaul latency can be developed, as well as suitable updates to the 3GPP timers can be done.
Figure 3-9: Intra AMF/UPF handover with Xn interface in place – [15], [16]

Tracking area Considerations:
The Tracking Area (TA) is the geographical combination of neighbour base station coverage footprints that is updated whenever a UE is camping (e.g. after handoff) in the coverage of a base station that was never visited before. Each Tracking Area has two main identities: Tracking Area Code (TAC) and the Tracking Area Identity (TAI). The concept of tracking area is particularly crucial for UEs that are either RRC Idle or RRC Inactive to reach them from the core network with Paging messages.

An extended radio access network can include cells that belong to different tracking areas, whereas each cell can only belong to one TA. When a mobile subscribes to the network, the AMF allocates a set (a “list”) of Tracking Area to the UE. All the tracking areas in a TA list to which a UE is registered are served by the same AMF. While a UE is in active state the UE location is known by the network at cell level. However, while the UE is in idle state its location is known by the network at TA level. Moreover, in case a mobile device remains inactive for a prolonged period of time the air interface connection and resources in the radio network are released. The mobile device is then free to roam between different base stations in the same TA without notifying the network to save battery and signalling overhead. In case a Paging command is sent to a UE, the AMF initiates the paging procedure by sending the Paging message to each gNB belonging to the tracking area(s) in which the UE is registered. Once the device responds to paging, the bearer(s) is (are) re-established.

If the UE detects that entered in a new TA which is not in the list of TAI that was included in the “attach accept” sent by the AMF at the time the UE registered with the network, as soon as the handover is completed, it triggers a Tracking Area Update (TAU) by sending a TAU request to the UE-serving AMF. This message includes a tracking area update for the last visited TAI and the UE ID previously allocated by the serving AMF. The new AMF selects a group of TAs to allocate and configures a new TAI list for the UE. Hence, a new Globally Unique Temporary Identifier (GUTI) may also be allocated.

3.4.2 Handover in legacy satellite system
In satellite systems, mobility was typically addressed by borrowing from terrestrial solutions like IP mobility, while at the same time striving to achieve handover independent mobility management, by binding addresses to geographical locations and ground segments. As shown in Figure 3-10 [17], beams are typically fixed and arranged considering the traffic demand, the satellite system
capabilities and the technology constraints. The mobility of a satellite terminal (e.g., in case of a moving platform) typically implies different actions for the satellite network management system depending on the nature of the handover.

- **Intra-satellite system handovers:**
  - Selection of the beam for areas covered by more than one (i.e., beams overlap),
  - Beam handover for satellite terminals, mainly based on location (geographical based beam handover),
  - Gateway handover, when the beams are not associated to the same gateway node (association of beams to the gateway is done at system design phase considering different parameters such as backbone capacity, security requirements, etc.).
  - Beam failover for satellite backup service (None GEO as main link and fallback to GEO)
  - Switching between Non-GEO and GEO based on various conditions (Load, Applications, Cost, Link condition etc.)

- **Inter-system handover:** handover between the satellite system and, typically, a terrestrial system.

As described before, handover in a satellite system can refer to terminal handover from one beam or transponder, gateway, satellite or terrestrial system, to another, implying some changes of the physical and logical resources, and significant requirements for synchronisation. For moving platform scenarios in which mobility takes place over geographical distances (aviation, maritime, etc.), beam handover is the most likely case to occur followed by gateway handover and then satellite handover, that may happen mainly in global scenarios, including inter-satellite systems (e.g., GEO-to-MEO).

As already mentioned, in a multi-beam gateway, depending on the location of the satellite terminal within coverage, a beam handover (if the terminal moves from one beam to another) may occur. Therefore, this implies that the satellite terminal should be tuned to the appropriate beam characteristics (frequency resources, access techniques, etc.) and it should configure all the required parameters for correct reception and transmission.

For service continuity, handover should also include terrestrial network handover, i.e. going from/to satellite network to/from terrestrial network with the basic principles described above that still hold. Such handovers may typically happen when the moving platform will reach its hub (train station, airport and port for respectively train, airplanes and ships) where connectivity typically is provided by a terrestrial system in addition/replacement to the satellite connection.

![Figure 3-10: Legacy satellite terminal mobility management based on geographical location](image)

### 3.4.3 Challenges of mobility management in SaT5G
Mobility management is one of the most crucial aspects that ought to be dealt with in a mobile network. Mobility management refers to the procedures that are set in place at protocol level to ensure that a user (a UE generally) can move from the coverage area of one base station (i.e., gNB) to another without incurring any service disruption. Handover is tackled considering the general case of a UE that roams around and performs a cell change. Specifically, in the SaT5G architecture implementation assumes that, except for the Layer 1 technology that might or might not be based on 5G NR, all other layers and procedures conform to 3GPP specifications.

For the sake of studying mobility, the different networking options that have been introduced in 5G, including relaying, have to be taken into account with respect to SaT5G options. For instance, it is worth mentioning the case in which gNB functionalities are delocalised as in a typical Central Unit (CU) - Distributed Unit (DU) functional split. The discussion on the different options will be resumed in Section 3.4.3.3 (see the scenario in Figure 3-15) and Section 3.4.3.4 (see Scenario Figure 3-17).

In the context of SaT5G, a taxonomy of the different architecture and implementation cases is shown in Table 3-7. Based on this, it is worth distinguishing the configurations in which the satellite segment is managed as a 5G compliant network from when it is not. Regarding this, SaT5G deliverable D3.1 [2] presented the possible different backhaul implementation options to a greater extent. At high level, the most crucial distinction consists of whether the satellite network conforms or not to 3GPP specifications, which implies also distinguishing the cases in which the relation between the terrestrial network and the satellite network can be considered trusted or untrusted in 3GPP terms.

Besides 5G NR direct access (out of scope in D3.2), the integrated 5G satellite-terrestrial network may identify the cases of a 'Transport Network'-based implementation and 'Relay Node (RN)'-based implementation. At a more granular level, the different backhaul implementation options described in [2] are the following:

- Indirect 5G UE access
  a) Transport network
    - Not based on 3GPP specifications
    - Based on 3GPP specifications
  b) Relay node
    - With untrusted non-3GPP access
    - With trusted non-3GPP access
    - With 3GPP access.

For the sake of mobility management in SaT5G, Table 3-7: Configurations for mobility management in SaT5G context provides an insight of the integration level between mobile and satellite networks in which each segment could either retain its own management system or this could be the same for both networks. The different configurations illustrate the concept: Indirect Access based on a relay node that, from an end-to-end perspective, includes 5G CN, the presence of a Donor node that through the satellite link connects to the remote relay and 5G UEs that serve users in the moving platform and the indirect access based on transport network that includes 5G CN, separate management for the transport, the satellite link with remote satellite modem, the gNB for access and the 5G UEs of the users. According to 3GPP Release 15 specifications, a relay node is part of the studies on Integrated Access Backhaul or IAB. Typically, the purpose of an RN is to provide range extension with respect to a parent base station, also referred to as the donor node.

Table 3-7: Configurations for mobility management in SaT5G context

<table>
<thead>
<tr>
<th>Configuration</th>
<th>Core Network</th>
<th>5G RAN</th>
<th>Remote area link</th>
<th>RAN</th>
<th>End user</th>
</tr>
</thead>
<tbody>
<tr>
<td>Indirect access based on RN</td>
<td>5G CN</td>
<td>Donor node</td>
<td>Satellite</td>
<td>RN (IAB)</td>
<td>5G UE</td>
</tr>
<tr>
<td>Indirect access based on TN</td>
<td>5G CN + TN</td>
<td>Satellite TN</td>
<td>gNB</td>
<td>5G UE</td>
<td></td>
</tr>
</tbody>
</table>
This section will concentrate on discussing the implications of mobility management procedures in the context of SaT5G, specifically the hard handover that has been considered from 4G onwards. Handover will be studied in the specific case of a moving platform (SaT5G UC4) that, while moving over geographical distances, can traverse different regions, thus changing satellite network coverage.

Although not explicitly evaluated in this document, which focuses more on architectural and protocol aspects, handover success rate and call drop rate are the typical KPIs that are measured by mobile network operators regarding the efficiency of the handover procedure.

### 3.4.3.1 Handover models in SaT5G implementation options

As already mentioned in [2], several different implementation options of the 5G integrated satellite network have been developed in SaT5G. In the 5G integrated satellite network, mobility management compliant with mobile network technology applies when the satellite system can be managed through 3GPP procedures, even in case the Layer-1 of the satellite is not supporting 5G NR technology.

**Assumption 1:** mobility management is addressed in a multi-beam satellite system that is typical in the next generation of Very High Throughput Satellite (VHTS) systems in which satellite beam coverage is either modelled as a cell sector or as full cell coverage in mobile network equivalents. Both cases of single gateway and multi-gateway will be analysed. It is also worth noticing that the description hereinafter is done mainly for the case of a bent-pipe satellite, since it is implicitly assumed that it will be the first steps integrated satellite-terrestrial 5G system.

**Assumption 2** transport network and relay implementation options are addressed. In the case of relay node implementation, 3GPP specification on IAB shall be considered [18]. Therefore, both CU-DU functional split cases and full gNB modes shall be analysed although not all the five relay options suit the present study. A discussion on IAB node aspects was undertaken already in Section 3.4.5 of [2] considering also that a dedicated UPF function can be collocated within the IAB Donor node.

**Assumption 3:** the study of mobility management relies on the hypothesis that the 5G integrated satellite system operates in 5G standalone (SA) mode with connection to the 5G CN. An example of study is shown in Figure 3-11: Example scenario for studying mobility management in case of an aircraft moving platform for the specific case of an aircraft moving platform, showing UP and CP protocol stacks, respectively.

**Goal of the study:** analyse existing procedures and develop those missing to assure service continuity in the 5G integrated satellite network in case of mobility in order to empower passengers of the moving platform with seamless 5G access.

**Analysis of pros and cons:** managing mobility in the satellite network in terms of 3GPP compliant procedures offers the implicit advantage of relying on well-defined and standardised protocols that shall allow using mobility management not only for service continuity but also to enable load balancing and intelligent traffic steering. These topics have been intensively researched in mobile network technology spurring the concept of self-organising network techniques for network self-planning and optimisation, which may become particularly relevant in future LEO satellite constellations. There is clearly an initial effort that is required to adapt the satellite system architecture to the procedures of mobile network technology. A specific problem is the satellite link delay, which ought to be taken into account during the effort to adapt mobile network techniques.

As already discussed extensively in [2], different satellite constellations impose very different delays on communication link, ranging from the 250 ms in GEO satellites (one-way) to nearly 40 ms and 8 ms for MEO and LEO satellites, respectively. With respect to this observation, it can be concluded that typical 3GPP timers for mobility management could not be re-used depending on the specific satellite constellation. Moreover, in case of GEO satellite, worldwide coverage is achieved with three satellites already and therefore, MEO and LEO definitively provide a more challenging environment regarding mobility management (LEO constellations might include thousands of satellites in the future). Further, IMT 2020 requirements prescribe control plane latency not larger than 20 ms for 5G [19], whereas 3GPP targets even the tighter value of 10 ms [20]. All these considerations are suitable inputs for standardisation impact toward the 5G integrated satellite network (e.g., 3GPP CT4 – Core Network and Terminals).
Mobility management in multi-satellite connectivity system:

The mobility management procedures identified in SaT5G implementation options rely on 3GPP defined mechanisms considering 5G SA radio connectivity to a 5G CN. In an advanced satellite system, the opportunity to switch connectivity between different constellations depending on the specific service that is dealt with (e.g. IoT scenarios) stems for an appealing feature to be added in the system. At the same time, mobility has to be efficiently managed also in this case.

Although not studied further in D3.2, the procedures developed hereinafter shall hold generally also in case of inter-satellite handover (e.g. GEO-to-MEO or MEO-to-LEO) with the main difference that different 5G core network functions may be involved, such as AMF re-selection. The inter-satellite system handoff may be modelled also as the handover that takes place between different mobile network technologies (e.g. 5G to 4G) and further adaptation of the procedures is likely to be needed.

Undesired effects in SaT5G mobility management:

In Section 3.4.1 it was already mentioned that the handover procedure may be subject to link failures that cause communication drop at the user side. In the context of SaT5G, in a moving platform such as a ship close to in-land mobile network base stations or on an airplane that is below three kilometres altitude, UEs may be affected by problems like the ping-pong effect and wrong handover. This could occur when the UEs start receiving signalling from other radio towers, which would not be the case for a ship far from the mainland or during the cruise time of a flight. Although such problems cannot be completely suppressed, they can be relieved by means of handover parameters optimisation for the case of a 5G network deployed on-board the moving platform.

3.4.3.2 Mobility management in 5G satellite system: the aircraft moving platform case

Relying on the description of handover procedures defined in 3GPP Release 15 that were provided in Section 3.4.1 and on the assumptions made in Section 3.4.3.1, relevant scenarios for developing mobility management procedures in the 5G satellite system are developed after selecting the important case of an aircraft moving platform (SaT5G UC4). The studies shall consider both single-gateway and multi-gateway scenarios with the due adaptations of the procedures that are presented in the following sub-sections.
Referring to Figure 3-12, the aircraft moving platform roams between adjacent beam coverages with each one that can carry different frequencies and cell IDs but in single-gateway case. The situation shown in Figure 3-12 closely resembles that incurred by mobile terminals that are connected to the terrestrial mobile network. Similarly, Figure 3-13 illustrates the case of an airplane that moves across different beams and satellite gateways, while maintaining connectivity to the satellite ground segment. Both options of single 5G core network and multiple 5G core networks are possible. Anyway, in the remainder of this study single 5G CN option will be further analysed. The main reason is to lay down the basic procedures for mobility management in the 5G integrated satellite network, considering that several different architecture alternatives are possible. Hence, the present study is tailored to the key principles of mobility management in relevant scenarios since they repeat in the different possible alternatives.

In addition, the new 5G integrated satellite system leverages the new concept of multi-criteria handover. This approach applies also to the case in which geographical areas are served by different satellite systems including GEO, MEO and low orbits satellites that may belong to different SNOs even. This abstraction is enabled by the adoption of SDN/NFV and coordinated orchestration of the integrated virtualised network, and the availability of enhanced interfaces at the control/management plane between satellite and terrestrial domains.

For completeness, the multi-criteria handover can target the following objectives although not restricted to:

- **Beam management**: used to re-direct the traffic in case of beam failure, overload or to anticipate such occurrences;
- **Beam handover**: when a terminal moves from one beam coverage to another (including update of associated gateway);
- **Beam handover based on service type**: the Network Management System (NMS) of the 5G integrated satellite system selects the beam in the satellite system (e.g. GEO or MEO/LEO) that meets at the best the requirements of a specific service;
- **Beam/Gateway handover based on traffic load**: the beam/Gateway handover can be performed to relieve the traffic load in the satellite network.
3.4.3.3 Mobility management procedures in single gateway multi-beam satellite

This section is devoted to study mobility management principles and procedures for the aircraft moving platform in a single-gateway multi-beam satellite system. The scenarios under study are shown in Figure 3-14 and Figure 3-15 in which it is assumed that the 3GPP NMS can either manage the 5G satellite network directly or exhibit some level of integration with the operator of the satellite system, which would depend also on the business model adopted. In both cases described by the figures, single 5G CN is in place to manage the 5G RAN on-board the aircraft. Further, the 5G CN falls within the domain managed by a MNO.

Referring to the Figure 3-14, for single-gateway multi-beam satellite configuration, the 5G UEs are located in the moving platform and connected to the gNB deployed on-board the aircraft. Through satellite TN implementation option between the satellite terminal on-board and the satellite gateway on the ground, the 5G RAN in the aircraft can connect to the 5G CN. When the aircraft moves around changing satellite beam coverage, typical beam management techniques adopted in satellite systems can be used with the techniques that change depending on the satellite constellation type.

Referring to Figure 3-15, for single-gateway multi-beam satellite configuration, mobility management is to be carried out in RN implementation option. As in the previous case, the 5G UEs in the aircraft are served by means of the 5G RAN on-board, which is exemplified by the IAB node. As previously described the IAB node implements 3GPP protocols and procedures and, through the satellite space segment connects to the satellite gateway on the ground. The IAB node implements two elements: the DU side that provides 5G connectivity to the 5G UEs and the Mobile Termination (MT) part that provides the DU in the IAB node backhaul connectivity to the CU in the IAB Donor node. In this specific case, it is assumed the CU-DU functional split option, but similar considerations apply to the case of full gNB implementation. Moreover, the CU of the IAB Donor is assumed located in the same infrastructure where the virtualised 5G CN is deployed but the same considerations hold in case the CU would be collocated with the DU in the gateway.

The 5G integrated satellite backhaul enables the connectivity of the 5G RAN on-board to the 5G CN on the ground with the MT that establishes RLC connection with the DU on the Donor node side. In this specific configuration, the IAB Donor node has one CU and a different DU in each satellite gateway (i.e., DU1 and DU2 in the example case of Figure 3-15). Since the CU on the Donor sides establishes RRC connectivity to the DU on the IAB node on-board the aircraft, when the airplane moves across different beam coverage areas, it will require only a UP adaptation that consists mainly of a data path switch (i.e., MT on the IAB node communicating with the proper DU in the IAB Donor), whereas the CP remains unchanged. In this RN case, satellite beam handover takes place with only PDU session reconfiguration at the user plane side from the 3GPP management perspective, thus not
requiring a full handover. The flowchart procedure shown in Figure 3-16 [18] was developed in the context of SaT5G using the 3GPP specifications available in [18]. To provide a concrete example, if a 5G UE on-board the moving platform is notified with a paging message from the 5G CN on the ground, the Donor IAB will send the message over the satellite link through the DU in the gateway and the satellite system that manages the beams will guide the paging message to the right beam where the aircraft is camping. When the user will move away from the airplane after landing, a legitimate handover procedure will be initiated toward the terrestrial network without undesired effects (e.g. wrong handover). In this way, the 5G UE will maintain connectivity from the time of departing to the time of arriving at destination except for the take-off and landing periods of the airplane.

Concluding, it is worth considering also that depending on the IAB architecture option [18], a UPF collocated within the IAB node can be deployed for the sake of managing the PDU session between the IAB Donor and dependent IAB node. This option is not analysed since no additional complication is envisaged for this IAB architecture option.

![Figure 3-14: Mobility management in 5G integrated satellite network in TN implementation option with single GW that does not require mobile network handover but satellite beam management](image-url)
Figure 3-15: Mobility management in 5G integrated satellite network in RN implementation option with single GW that does require handover in multiple gateways case with single CU and multi-DU.

Figure 3-16: Path switch selection in single GW single CU multi-DU IAB satellite with mobile relay.
3.4.3.4 Mobility management procedures in multi-gateway multi-beam satellite

In multi-gateway multi-beam satellite configuration, the aircraft moving platform will pass from one satellite beam connected to a gateway to another beam connected to another ground gateway station (in principle they can be located at the same physical gateway site or in a completely different location). The specific scenario for mobility management is shown in Figure 3-17 in which satellite gateways are assumed physically separated. The 3GPP management system can manage directly the 5G satellite network, while the management of the space segment is left to the SNO. Similar to the case analysed in Section 3.4.3.3, the CU-DU functional split is further assumed. The system shown in the figure stands for an RN node SaT5G implementation option and also in this case the specific business model adopted allows different levels of cooperation between MNO and SNO.

Each IAB Donor node has its own separate CU that is assumed to reside in the infrastructure where the virtualised 5G CN managed by MNO is deployed. On the other hand, a separate DU is hosted in each gateway station that connects to the corresponding CU. Referring to Figure 3-17, the pairs are CU1-DU1 and CU2-DU2. The 5G UEs on-board the aircraft are connected to the 5G CN on the ground whereby the backhaul connection established through the IAB node in the aircraft. The DU part of the IAB node provides user plane functionalities to the 5G UEs and, through the MT part of the IAB node connected to the DU part of the IAB Donor (e.g. DU1), control plane association is maintained with the CU in the serving IAB Donor node (e.g. CU1). Following the motion of the aircraft, when it reaches the overlapping area between adjacent beams that connect to different gateways, a full 3GPP handover has to be carried out. For completeness, in order to simply the procedure, an Xn interface is assumed between different Donor IAB nodes that in Figure 3-17 are denoted with coverage areas of different colours. The Donor IAB node that currently serves the IAB node on-board the aircraft is denoted as Source IAB Donor, whereas the donor to which the IAB node is relocated as Target IAB node. It is also worth reminding that in case the beams that are traversed by the moving platform are served by the same CU entity, the procedure analysed in Section 3.4.3.3 can be reused.

For completeness, a separate UPF can be co-located with each CU to manage the PDU session between the 5G UEs and the 5G CN on the ground. As in cases previously discussed, the CU of the IAB Donor node can be also co-located with the DU in the satellite gateway. Moreover, referring to the IAB node architecture options detailed in [18], similar reasoning applies in case full gNB would be considered instead of the CU-DU functional split case.

The handover procedure developed for this case of RN mobility management is shown in Figure 3-18. The mechanism relies on the 3GPP procedures defined in the Release 15, but it is worth stressing the fact that mobility management in relay node configurations in which the mobile relay changes association to the IAB Donor is currently not addressed by existing 3GPP specifications and it was developed explicitly for the sake of SaT5G moving platform scenarios. Clearly, a full 3GPP handover in the context of a 5G satellite network is not trivial and handover failure and connection drop rate should be evaluated. In any case, optimisation of handover parameters and self-organising network techniques are extremely important and pave the way to future studies.

As a concluding remark of this section devoted to mobility management aspects in SaT5G moving platform scenarios, in case the IAB node on-board the aircraft is simply de-attached from the Source IAB Donor and it re-attaches to the Target IAB Donor, service continuity cannot be preserved for the 5G UEs of the passengers, thus emphasizing the importance of the mobility management procedures.
Figure 3-17: Mobility management in 5G integrated satellite network with multiple GWs and CUs

Figure 3-18: IAB node handover procedure in RN when changing IAB Donor node in multi-GWs, multi-beam 5G integrated satellite system and IAB architecture implements CU-DU functional split
3.5 **Multilink management**

This includes multi-satellite connectivity and dual satellite-terrestrial connectivity between satellite and terrestrial networks. Multi-satellite connectivity refers to access network being connected to the core network through two or more satellites at the same time. Dual satellite-terrestrial connectivity refers to an access network being connected to the core network through a satellite link and a terrestrial link at the same time. The latter case has a parallel in the concept of dual terrestrial connectivity for 4G and 5G, in which for example a reliable but low-bandwidth 4G channel is used for the control plane, and a high-bandwidth but less reliable 5G channel is used for the data plane.

The 5G system should be able to take advantage of these backhauling options in a way that improves the user experience, optimises the traffic distribution and enables the provision of new high-data-rate services.

For example, the scenario envisaged for dual satellite-terrestrial connectivity could be to carry the control plane on the terrestrial direct ground 4G link, and the user plane on the satellite link.

![Figure 3-19: Multi-satellite connectivity](image)

The techniques to perform the traffic steering, splitting and switching (TSSS or T3S) will depend on the considered implementation options (RN or TN).

Typically the TSSS for RN implementation will be based on existing ATSSS specification proposed in 3GPP while for the TN implementation options, the development of a specific hybrid multiplay function (HMF) is foreseen in the frame the Satellite core network.

These solutions are further described in sections 3.5.1 and 3.5.3.

### 3.5.1 Protocols for traffic steering, splitting and switching distribution

TSSS investigated in this document refers to the TSSS at the backhaul level. For RN based implementation options as well as TN-based implementation options, a key challenge is to select the appropriate protocol to perform the traffic distribution.

When the TSSS is performed using criteria available at RN or gNB level, it can be easy to select a given link for certain type of traffic (e.g. control plane via terrestrial, user plane via satellite).

The access point can easily split these two types of traffic since it is already visible or even initiated at this level.

When the criteria of TSSS are based on the UE traffic flows or based on UE characteristics, the multilink functions need to be able to detect these criteria and also to split or switch a specific flow and recombined it at the other side of the link, before delivery to the destination.
Among different protocols that can be used to perform the TSSS, 2 options can be envisaged for the TSSS at the backhaul level, whatever the implementation option is: MPTCP and MPQUIC.

- **Distribution via MPTCP:**

An MPTCP (RFC 6824) connection begins similarly to a regular TCP connection. It can be genuinely integrated in RN functionalities for RN based implementation options and for TN based implementation options, extended APIs inserted in HMF function could provide additional control to MPTCP-aware applications.

MPTCP can identify multiple paths by the presence of multiple addresses. When extra paths are available, additional TCP sessions (termed MPTCP "subflows") are created according to networks characteristics, and are combined with the existing sessions, which continue to appear as a single connection to the applications at both ends.

![Figure 3-20: Example MPTCP Usage Scenario in HMF implementation](image)

The MPTCP function has to keep track of TCP segments, and present them to the session in the correct order. If the MPTCP subflows have different latencies, then the flows with the lower latencies must be buffered, and if one subflow is over a GEO satellite link, then the other flows must contain sufficiently large buffers and the overall performance of the MPTCP is limited by the flow with the greatest latency [21].

- **Distribution via MPQUIC:**

QUIC is an experimental transport protocol developed by Google since 2012. It improves the perceived performance of connection-oriented web applications while requiring little or no change from applications. QUIC is very similar to TCP+TLS+HTTP2, i.e. leveraging HTTP/2's multiplexed connections, allowing multiple streams of data to reach all the endpoints independently, but implemented on top of UDP.

Key advantages of QUIC over the TCP+TLS+HTTP2 combination:

- Connection establishment latency;
- Improved congestion control;
- Multiplexing without head-of-line blocking;
- Forward error correction;
- Connection migration;
- Connection Establishment.

QUIC handshakes frequently require zero roundtrips before sending payload, as compared to 1-3 roundtrips for TCP+TLS. This is a strong advantage in high latency environment where it would dramatically reduce the session establishment. Since QUIC is working over UDP, the protocol does not force in-order delivery of packets thus QUIC avoids HOL blocking. QUIC connections are identified by a 64 bit connection ID, randomly generated by the client. When a QUIC client changes IP addresses, it can continue to use the old connection ID from the new IP address without interrupting any in-flight requests. This connection migration feature is interesting when a user migrates between two links, such 5G to Wifi or satellite. Finally, QUIC runs in application space instead of the kernel space allowing faster improvements.

In June 2015, QUIC was submitted to the IETF, QUICK Working Group was established in 2016 and a request was made in October 2018 by the HTTP and QUIC Working Groups to rename the protocol HTTP/3 to prepare a standard.
The study “How quick is QUIC?” [22] shows that QUIC outperforms SPDY and HTTP protocols in lossy environment: QUIC is the fastest protocol being roughly 25-30% faster than SPDY and 35-40% faster than HTTP.

In the context of the SaT5G project, it is important to consider that QUIC is designed to prevent middle boxes to tamper with the transported content, and to protect its connection from various attacks. It is not possible in particular, to forge or modify Ack messages. This leaves little room for PEPs dealing with QUIC. However, it is possible to force end to end QUIC connection attempts to go back to TCP transport. This feature was introduced as a mean to allow connection through firewalls.

Following their work on MPTCP, the University of Louvain has published Multipath QUIC as an extension to the QUIC protocol that enables hosts to exchange data over multiple networks over a single connection [23]. Bringing Multipath to QUIC is relatively simple and it could provide a better integration with QUIC than TCP.

These aspects are further investigated in WP4.3 – Multi-link and heterogenous transport.

### 3.5.2 Relay-node implementation

The AT3S functions in the 5GC and the UE for the TSSS mechanisms can be implemented at the relay node level in order to provide this TSSS at the backhaul level, rather than the UE access level.

An adaptation of UE-AT3SF [24] would therefore being performed for the RN node (named RN-AT3SF) and updated at the core network side to support TSSS at RN level. According to the information received from the core network over N1 interface from PCF or SMF, the Relay Node would be able to steers, switches or splits the flows in order to manage the traffic between the available links (e.g. GEO or MEO or LEO satellite link, point-to-point RF, terrestrial link, etc.) in a way that is transparent to the user.
As for the UE, traffic steering control would be triggered by the PCF initiated request and consists of steering the detected service data flows matching application detection filters or service data flow filter(s) in PCC Rules. Upon receiving a PCC rule which contains the traffic steering control information, the SMF would provide the information to the UPF for the enforcement.

As for the UE again, traffic splitting would be based on a route selection descriptor sent by the network (PCF), user preference, and link performance. The ATSSS would provide Service Data Flow (SDF) level steering capability to establish new PDU session(s) over the preferred network based on traffic descriptor and rule precedence.

Measurement reports may be delivered, on request from the network (UPc-AT3SF) when certain threshold levels for the value of the measurements are reached.

The user reporting rules from both links (satellite and non-satellite) are sent to the SMF which will modify the steering, switching and splitting configuration in accordance with others parameters like T3S policies and UE profiles and subscription.

These concepts will need some standards evolution, typically in 3GPP SA2 working group, in particular in order to allow the support of the TSS at the RN level.

### 3.5.3 Transport network implementation

In the transport network option, a Hybrid Multiplay Function (HMF) is logically implemented between the RAN and the SAT Terminal, (physically, it can be incorporated in the same equipment) at the edge side and between the SAT GW and the core network as shown on figure 32.
Since the multiple links are available at the backhaul level, instead of Access Traffic Steering, Switching and Splitting (AT3S) as designed in 3GPP systems, a proposed function named Hybrid Multiplay Function (HMF) between different satellite links or satellite link and a non-satellite link is foreseen to perform the TSSS.

The traffic can be split over different links for equal or weighted distribution, steered to the least cost or the best performing network or switched from one link to another if a link condition falls below a given performance thresholds based on for example link quality, throughput, latency, packet loss, etc.

The Network Management should be able to provide to the HMF information from data flows (e.g. QoS rules) and network characteristics (delay, performance, congestion, etc.) for the management of the appropriate exploitation of available links. A close coordination between the satellite TN NMS and the terrestrial 3GPP NMS will be necessary to allow such level of interaction.

### 3.6 Slicing and virtualisation

The network slicing is a key enabler for the future 5G networks. 5G systems will enable isolated logical network slices (over a common physical infrastructure) to serve several tenants and enabling service-based networking. Each network slice will involve several domains and imply a cross-domain resource slicing handling.

In the frame of SaT5G, the above-mentioned domains are identified as:

- 5G Core network;
- Satellite network;
- RAN;
- UE.

Figure 3-25 depicts these domains which have to support the network slicing from SaT5G perspective. Among all other possible options, SaT5G introduces a satellite link between the 5G CN and the RAN; satellite has no option but to support the network slicing in order to handle an E2E slice from the UE to the 5G CN. Embracing SDN/NFV paradigms in satellite systems will be key assets to provide such support. These aspects are introduced in terms of requirements for network function virtualisation in section 3.6.2 and further investigated WP4.1 and WP4.2.

![Figure 3-25: Cross-domain slice](image)

This section aims to propose a definition of the slice criteria and the requirement for network function virtualisation to support network slicing.

#### 3.6.1 Management of network slicing and slice definition criteria

The cross-domain management of network slicing is a key challenge for the 5G satellite and terrestrial integration and need to be supported by the different domains: terrestrial domain and the satellite domain. Depending on the backhaul implementation option, the management of the different domains can be performed following two approaches:

- **Single 3GPP Network Management System (NMS)** for the relay node based implementation options. As depicted in the Figure 3-26, the 3GPP NMS will simultaneously manage the terrestrial domain (UE, RN, 5GCN) and the satellite domain. The requirements of a specific slice will therefore be directly endorsed by the single 3GPP NMS which will derive the required configuration for the components in both terrestrial and satellite domains and manage the end-to-end slice configuration;
**Dual NMS:** 3GPP NMS and SatCom NMS for the transport network based implementation options, with the satellite domain being managed by the SatCom NMS and providing advanced features to the 3GPP NMS such as support of Network slicing. The slices requirements and other requests can therefore be sent by the 3GPP NMS to the SatCom NMS by the coordination interface, and the Satcom NMS will setup the satellite system following the slice requirements provided by 3GPP management.

These ways of network management lead in any case to the definition of slice expressed by means of the slice requirements.

It is worth noting that here we assume that there is no limitation from the infrastructure or device perspective to support different validated slice requirements and we just focus on the definition of the slicing concept in the integrated satellite terrestrial networks.

Some slice definition criteria are provided below:

- **Use case-based slice:**
  An end-to-end network slice can be based on use cases and is then dedicated to one use case. These use cases are typically the IMT-2020 usage scenarios, also known as eMBB, uRLLC and mMTC. These three major use cases have a broad variety of capabilities and hence are one of
the most rational ways to define a network slice. In addition, SaT5G use cases defined in [3] may also tackle this role of slice definition. All these use cases can possibly lead to distinct technical requirements and then are considered as network slicing criteria.

- **Application-based slice:**
The applicative criterion enables a slice definition as an application type dedicated slice. A type of application is defined by a set of parameters such as bandwidth, delay, resilience or security. A combination of specific settings of these parameters leads to an application type definition. Therefore, with this logic, we could consider that any service could be mapped on one specific type of slice.

- **Tenant:**
The tenant criterion enables a slice definition as a stakeholder dedicated slice. This concept allows a SNO or a MNO to create an end-to-end network slice for a third party like e.g. a MVNO. This slice is then devoted to the MVNO utilisation which implies a high-isolation customisation.

The specification of these slices criteria and their associated requirements can be realised in two manners: following a top-down approach or bottom-up approach:

- For the top-down approach, the slices are defined within the terrestrial network and the satellite network is asked to support them. This top-down approach is a long term vision of slice definition, because all the domains have to support the slicing and coordinate each other. The satellite network will have to implement efficient techniques and provide enhanced interfaces in order to support such request. However, the defined slices can still be based on criteria defined in a Service Level Agreement (SLA) between the different domains;

- The bottom-up approach means that the slice definition criteria are mostly provided by the satellite domain. The approach consists on exposing which type of slice the SNO could support. This slice would be tailored to the SNO resources and would be typically exposed as an applicative slice. The criteria considered for defining the slice can be any of the criteria presented above (i.e. based on use case, application, tenant, etc.). In this approach, SatCom systems could define a restricted number of slices which they can support, based on their infrastructure and management system capabilities. Then, the satellite management system could expose its available (i.e. supported) slices to external parties (e.g. MNO). A third party will therefore be aware of the slicing capability of the satellite link and will set up a consistent end-to-end slice. This may be a short or mid-term solution because it could rely on the legacy satellite systems. Furthermore, another side of the bottom-up approach could be to propose a mapping of the usual SNO slices on the MNO slicing definition following the bottom-up approach.

### 3.6.2 Network Function Virtualisation requirements for network slicing

The promoted adoption of network slicing over various types of networks implies to develop efficient ways to perform slicing. In this context, this section focuses on deriving the NFV requirements to support the previously defined slicing.

The challenging integration of the satellite into the future 5G networks will likely relies on a wide adoption of the virtualisation. This approach will provides the satellite system with a high level integration in terms of flexibility, network management or function exposition. For better integration, the satellite system has to virtualise its own network functions. This will lead to global management perform by an orchestrator which will be able to control and integrate the satellite system.

In this perspective, as many satellite functions as possible have to be virtualised. All the protocol stack of OSI layers can be affected, even L1 layer considering SDR paradigm. Then, NFV requirements in terms of CPU, storage or network resources can be derived after identifying the lowest OSI layer at which the desired function will operate. This level is the definition of the virtualisation depth which characterises the NFV requirements for a specific function.

Moreover, a certain level of virtualisation is needed in all the four domains (5GC, satellite system, RAN and UE) in order to consistently manage an end-to-end slice. That means, in addition to the 5GC which is developed to support SDN/NFV and slicing, RAN, UE and of course satellite systems have also to handle SDN/NFV and slicing requirements.

### 3.7 Features related to the implementation options
This section deals with the distinctions between the integration of the new features according to the implementation options. The aim is to provide first requirements for the five indirect access implementation options defined in the frame of WP3.1.

The section begins with the transport network options:

- Transport network non-based on 3GPP system specifications;
- Transport network based on 3GPP system specifications.

The analysis continues with the relay node options:

- Relay node based on untrusted non-3GPP access;
- Relay node based on trusted non-3GPP access;
- Relay node based on 3GPP access.

### 3.7.1 Transport network non-based on 3GPP system specifications.

This transport network implementation option introduces an adaptation layer which wraps the non-terrestrial network and it participates at several levels.

The adaptation layer has to interconnect the 3GPP management system and the transport network management system. This link is a key enabler for satellite-terrestrial integration. Moreover, this specific implementation could be a short term option which could be adapted to the legacy systems whether they are terrestrial or satellite. All the network convergence work would be handled by this adaptation layer.

From the management perspective, the adaptation layer has to fulfil the requirements of network function exposition. This enables a better interfacing between TN management system and 3GPP management system. The adaptation layer shall provide a bridge role at control plane level. Depending on the integrated features, this layer has to manage the following points:

- **HMF control plane protocol stack**: the adaptation layer have to handle the adequate TSS protocol and all the related requirements;
- **QoS adaptation**: if the mechanism is based on RRC signalling coming from AMF and SMF, the mapping between 5QI and sat CoS and then sat resources would probably be performed by this layer;
- **Level of NFV**: based on the infrastructure flexibility, the virtualisation depth available has to be exposed by the adaptation layer.

![Figure 3-28: Transport network non-based on 3GPP system specifications](image-url)
The 5QF and SatQF functions are respectively the 5QI QoS Function which performs the mapping between the 5QI and the satellite CoS, and the Sat QoS Function which performs the mapping between the satellite CoS and the satellite resources. These functions would likely be integrated at NTN level.

### 3.7.2 Transport network based on 3GPP system specifications

This transport network implementation option is based on 3GPP system specifications. That means TN management system has to provide 3GPP management system with an exhaustive exposition of the satellite specific network functions supported. In particular, TN would have to expose its satellite CoS in order to enable RAN to perform the 5QI to satellite CoS mapping.

![Figure 3-29: Transport network based on 3GPP system specifications](image)

### 3.7.3 Relay node based on untrusted non-3GPP access

Untrusted non-3GPP accesses are in the scope of the 3GPP R-15. Basically, the UE gets an IP address from the N3IWF function and initiates an IPsec Security Association (SA) in order to establish an IPsec tunnel between the UE and the N3IWF function. The UE is then connected to the 5G core through N3IWF which itself is connected via N2 and N3 interfaces (respectively to the control plane (AMF) and user plane (UPF)).

The multilink support is provided by means of two functions. First of all, AT3S function enables traffic steering, switching and splitting between a 3GPP access and an untrusted non-3GPP access. Subsequently, this relay node based on untrusted non-3GPP access has to handle the multilink by adapting the RN as a RN-AT3SF.

The QoS adaptation support can be provided by the NTN terminal. In this implementation option which relies on a unique 3GPP management system, the mapping between 5QI and satellite resources, performed by the 5QI to Satellite resources QoS Function (5sQF), could be managed by NTN terminal. Mapping table would be implemented and updated by control plane signalling message (e.g. RRC). Moreover, N3IWF should determine the number of IPsec child SA to create and associate QoS profile to each IPsec child SA.

The network slicing options depend on the flexibility provided by the NTN terminal. If an L2 or even deeper NFV is targeted, the main issue will probably raise from the NTN terminal and its virtualisation capacity.
3.7.4 Relay node based on trusted non-3GPP access

Trusted non-3GPP accesses are in the scope of the 3GPP R-16 and currently being studied. In this implementation option, the access is provided by a Trusted Non-3GPP Access Network (TNAN) which does not employ a 3GPP standardised technology and can connect to 5G core via N2 and N3 (i.e. compliant interfaces).

From a network function perspective, TNAN can be divided in two network functions: A Trusted Non-3GPP Access Point (TNAP) function which enables the access of the UE to the TNAN; A trusted Non-3GPP Gateway (TNGF) function which exposes the N2 and N3 interfaces in order to connect the UE to the 5G core. This structure is depicted on Figure 3-31. For the sake of completeness, it is worth to note that a TNAN can be composed of multiple TNAP linked to one or more TNGF respectively. If there are multiple TNGF, there is a link between each TNGF and 5GC (i.e. N2 and N3 interfaces). Optionally, an inter-TNGF link may be supported; this architecture is depicted on Figure 3-32.

Figure 3-31: Structure of TNAN [25]
Figure 3-33 depicts the suggested SaT5G architecture for supporting the new features. As for the untrusted non-3GPP access, the multilink is managed by an adapted relay node (i.e. RN-AT3SF). The QoS adaptation should be provided by the NTN terminal and the NTN GW. These are the two functional parts of the TNAN. Nevertheless, functionally speaking, TNGF would probably be integrated to the RAN Donor node because of its 3GPP affiliation. Moreover, the slicing support would be the same as the untrusted non-3GPP access (3.7.3).

3.7.5 Relay node based on 3GPP access

In this implementation option, the RN is connected to the Donor and so to the 5GC through 3GPP access network. 3GPP management system is then fully aware of the whole integrated network which can support all the new features. The multilink is provided by a RN-AT3S, the QoS is adapted by the NTN terminal or the NTN GW but both RN and Donor node are aware of the QoS capability of the satellite link.

This implementation would probably be the long term option and it would be particularly well suited to handle a really deep NFV (e.g. L1). Everything would be managed by the 3GPP management system and the fully integrated network would facilitate the mobility management.
Figure 3-34: Relay node based on 3GPP access
4 Network deployment analysis

4.1 Considered role model and backhaul implementation options

4.1.1 Role model

SaT5G deliverable D2.3 [13] addresses the business modelling of integrated 5G terrestrial and satellite networks. The actors and a generic value network for the 5G are proposed. The role of a broker to manage the supply and demand risks for both MNOs and SNOs has been defined. The broker finds different offers in the satellite market and deals with satellite network operators (SNO) or with others brokers (which might be developed by SNO or third parties). A broker carries the risk of dynamic spectrum allocation and it must have the ability to manage heterogeneous satellite systems.

In this broker concept, it has been identified a number of physical and technical constraints that make sharing resources coming from different SNOs a complex endeavour. A subset of these issues is introduced below with potential solutions to enable the proposed broker role:

- Need to repoint between GEO HTS to change SNO (motorised reflectors and flat panel arrays can address this);
- Different SNOs use different vendor systems (eventually a standardised based approach using mostly VNFs may address this);
- Different SNOs use different frequencies (most GEO use Ka band however Ku band is also used, nGSO use both Ka and Ku bands);
- Antennas and systems optimised for GEO operation will not readily work with nGSO and vice versa (flat panel arrays, software defined radios and standardised based modems using mostly VNFs may address this).

The applicability of an automated process is considered, especially when a time-consuming manual process is required.

A variety of relationship models are considered and the selected role model represented in Figure 4-1 below, assumes a single SNO providing service to several MNOs. Other SNOs can also have similar relationships with a potentially overlapping group of MNOs.

![Figure 4-1: Role model 2 (Several MNOs sharing Satellite resources)](image-url)
This is the most likely role model for the **short to medium term** given the technical constraints and investment cycles.

Figure 4-2 shows the proposed design approach to the broker concept. For the sake of completeness, it is useful to mention that previous research projects like the H2020 VITAL [26], the H2020 COHERENT [27] and the H2020 5G Exchange [28] studied concepts related to resource brokering analysed from different perspectives. The VITAL project took the approach of developing a federator of resources; COHERENT proposed a way to use abstraction models to control and manage network resources; and 5G Exchange delved into the design of multi-domain orchestration. In the development of the broker concept, SaT5G will benefit of such previous work, elaborating the fact that the broker provides mediation at business level between different parties that are on one hand offering resources (i.e. sellers) and on the other those that are seeking resources (i.e. requesters). Figure 4-2 embraces this approach, with customers’ entities that send their service requests to the broker. Specifically, the broker is devised as a software entity trusted by all parties involved in a transaction. Although in the context of SaT5G customers are mainly operators, also third parties can exploit the presence of the broker, which would be responsible of trading the resources. From the logical standpoint, the broker is the highest point in a hierarchical model that exposes bundles of resources leveraging on abstraction models. The resources exposed to the Market Place may belong to a single technology domain, as well as group of resources may be aggregated from multiple domains (i.e. SNOs in SaT5G). In the specific case of SaT5G, a technology domain is that of an SNO. As shown in the figure, domain specific orchestrators and an optional multi-domain orchestrator (concept investigated [28] and [29]) are meant to instantiate network services and manage virtualised resources end-to-end, as well as manage the lifecycle of the instantiated services. Moreover, the broker shall maintain repositories of virtual and physical resource availability in each operator’s domain in the form of databases, which are thus exposed through the abstraction models already mentioned. The Mediation Layer should stand for the policy used to assign the resources traded between requesters and sellers. The policy can include different mechanisms including bidding schemes, optimisation techniques, game theoretic methods and so on.

From the perspective of the Management and Orchestration layered architecture, the broker would endorse the role of a federator, which is able to interact with different domains orchestrators through their NorthBound APIs.

![Broker entity layered view representation](image)

Figure 4-3 shows the workflow that is envisaged to take place between resource requester(s) and seller(s) during the trading process mediated by the broker entity. Here, the workflow is discussed from the perspective of the broker assuming, in case of SaT5G, that requesters are MNOs or third parties (e.g. an airline in case of SaT5G use case 4). As mentioned, in the context of SaT5G, the pool of resources is mainly satellite bandwidth made available in the market by SNOs, which act as sellers.
Advanced concepts can include the MNOs as resources sellers, typically when the requester is a third party (e.g. an airline) without particular engagement with a specific MNO, and warranting a technical ability to connect to any of them.

At the broker level, the satellite resources are exposed to the requester(s) through abstraction models, thus exposing bundles that can be specified with attributes like bandwidth, delay and satellite type, to mention a few. As discussed previously, the satellite bandwidth may be the composition of the resources that belong to different SNOs and the aggregation over different satellite constellations.

The negotiation process arbitrated by the broker begins with the preliminary phase in which a resource availability check is carried out. The format of the request, whose definition is out of scope of this work, should allow describing attributes like bandwidth, latency, preferred satellite type (e.g. GEO), etc. The attributes specified in the request might be binding or just set preferences. In addition, geographical location information where the service has to be enabled should also be provided to the broker. Upon receiving a request specified as mentioned, the broker first checks whether a report was already generated for that specific request or not. If this is either a new request or the resubmission of a rejected one (modified with respect to the original), the broker will proceed with preliminary verification of resource availability.

As a subsequent step, the broker retrieves resource availability from the internal repositories spanning across all technology domains that are subscribed to that specific broker. Clearly, it is crucial that the information stored in the repositories are updated as frequent as possible so that decisions are not made on outdated information. The broker will validate the request verifying whether the type of resources is part of the current offer portfolio and that those are supported in a specific geographical region. In positive case, as the next step, the broker verifies that there are enough available resources to satisfy the request. If this preliminary phase of the resource negotiation is successful, the broker shall generate a report in response to the request. If the requested type of resources is not supported, the broker can make an attempt to retrieve resource availability from other satellite systems that can still satisfy the request, in case the attributes are set as preferences and thus are not binding. If a different type of resources is supported and available, the broker will respond with a report also in this case. On the other hand, if the verification phase of resource availability fails, a rejection shall follow. A rejection as response from the broker might trigger the resubmission of a new modified request.

This role model and its associated workflow are therefore considered in this document for the design of a network deployment which meets the requirements of a SaT5G scenario. SaT5G scenario 2b and scenario 4a are the two scenarios selected for the analysis.
4.1.2 Considered backhaul implementation options

Considering the Table 2-1, defined in D3.1 [2], which provides a comparison of the satellite backhaul implementation options in terms of key challenges and time to market, it can be derived that for the **short to medium term** time to market targeted by the proposed role model two backhaul implementation options can be considered in this analysis:

- RN based on untrusted non-3GPP access: challenges such as implemented untrusted access mechanisms as requested by 5G standards and adaptation of relay node mechanisms to satellite terminal need to faced;
- TN based on 3GPP system specifications: designing a specific “5G ready” satellite transport network based on 5G system specifications will be the main challenge to face, with the development of advance interface at the management level in order to allow a close coordination between the satellite NMS and the terrestrial 3GPP NMS.
4.2 Network design for SaT5G scenario 2b

4.2.1 Scenario requirements

As defined in D3.1 [2], the **SaT5G scenario 2b – satellite backhaul to individual cell towers**—refers to satellite backhaul to individual cell towers. The description of the scenario is reminded in the following box:

**Satellite backhaul Scenario 2b:** Satellite backhaul to a single cell tower located in a rural area in the EU covering two villages about 5km apart and a rural main road.

The villages are home to 300 families, in summer months an additional 50 families may be in holiday accommodation. The road can occasionally be busy with holiday traffic but is usually quiet. The predominant traffic on the cell is eMBB but there is some mMTC traffic generated by agri-tech.

An application case for this scenario is proposed in 3GPP TS 22.822 [30] as follows:

**Two small villages (Aville and Bville) in the country about 5km apart are connected by a main road. The two villages do not currently enjoy modern communication services. DSL connectivity is poor because of the distance to the next small town, Cville, where the mobile operators also have good coverage, but it hardly reaches the two villages. The mobile operators have decided to place a shared cell tower between the villages. A satellite operator has excellent coverage of Aville and Bville. The mobile operators have decided to start up by using a satellite backhaul to the new cell tower. This can be arranged quickly and there is no need for any construction work, beyond the site of the cell tower. The MNO centralized Resource manager (RM) realized that an area with population is not covered by the network. It triggers an alert to the operator business centre that approves a new infrastructure deployment. The deployment in this case is a single cell tower located in a rural area.**

4.2.2 Negotiation scheme between the actors for the SaT5G scenario 2b

Following the role model described in section 4.1, and inspired from the example of workflow represented in Figure 4-3, a negotiation process can be initiated by the broker with different SNOs to address a request from a MNO that is interested by deploying a network to target the considered SaT5G scenario 2b. At the end of each process, the broker sends the set of results in a report to the MNO. The report includes different options:

- Fixed amount of satellite bandwidth with a fixed price per month;
- A minimum bandwidth granted with a fixed price plus fair dynamic bandwidth allocation in a model pay-as-you-go;
- A minimum bandwidth granted with a fixed price plus premium dynamic bandwidth allocation where it is possible to pay more to use the dynamic spectrum;
- Fully dynamic where resources are not granted;
- Any other options.

The MNO finally decides which resource option is best. Of course, in order to use resources from another provider, the two stakeholders must have a pre-established Service Level Agreement (SLA).

4.2.3 Integrated backhaul architecture for SaT5G scenario 2b

Considering the satellite backhaul implementation options and the requirements associated to the SaT5G scenario 2b, the following Table 4-1 identifies the required high level functionalities to be implemented in the integrated network. As stated in section 4.1, the considered backhaul...
implementation options are RN based on untrusted non-3GPP access and TN based on 3GPP system specifications since the term targeted is short to medium.

**Table 4-1: SaT5G scenario 2b analysis**

<table>
<thead>
<tr>
<th>SaT5G High level functionality</th>
<th>Scenario requirement</th>
<th>Role model requirement</th>
<th>RN based Implementation option requirement</th>
<th>TN based Implementation option requirement</th>
</tr>
</thead>
<tbody>
<tr>
<td>Registration / Connexion</td>
<td>No</td>
<td>Yes (for accounting and billing)</td>
<td>Yes (since RN managed by the 3GPP NMS)</td>
<td>No</td>
</tr>
<tr>
<td>gNB/RN roaming</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Resource allocation</td>
<td>Yes (to meet the requirements of data rate/user)</td>
<td>No/yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>QoS adaptation</td>
<td>Yes (to meet the services QoS requirements)</td>
<td>No</td>
<td>Yes (mapping to satellite resources)</td>
<td>Yes (mapping to satellite CoS and then to satellite resources)</td>
</tr>
<tr>
<td>RAN aspect (incl. mobility management and beam handover)</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Multilink management</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Slicing support</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Support of MEC</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
</tbody>
</table>

Consideration of this table can lead to a selection of the features and the architecture of the designed backhaul network.

Figure 4-4 provides the detailed backhaul architecture for SaT5G scenario 2b with RN implementation option based on untrusted non-3GPP access.
The Figure 4-5 provides the detailed backhaul architecture for SaT5G scenario 2b with TN implementation option based on 3GPP system specifications.

![Figure 4-5: Backhaul architecture for SaT5G scenario 2b with TN implementation option](image)

### 4.3 Network design for SaT5G scenario 4a

#### 4.3.1 Scenario requirements

As defined in D3.1 [2], the SaT5G scenario 4a – Updating content for on-board systems and grouped media request by the moving platform company – refers to the update of media content for on-board systems and grouped media request by the moving platform company. The description of the scenario is reminded in the following box:

This scenario can be resumed as a grouped request for media by the moving platform company.

This would be a way to update the content proposed by the moving platform company to passengers and subscription to live TV. The end user would use a standalone application on its own device or application pre-installed on the devices provided by the company.

The catalogue is updated with predictive valuable content and most demanded content. Accessible media might include videos, music, game patch and newspapers. Live TV can also be proposed to broadcast for example live TV Show, TV News or Live Sport program like a champion’s league game.

This Scenario is tightly related to caching/multicast edge delivery, exploiting the inherent broadcast capabilities of satellite networks and adding the particularity of being addressed to moving platforms. It is considered of high added-value for the moving platform companies being potentially combined with a full broadband access (as described in Scenario 4b).

A stand-alone satellite backhauling can be envisaged for airplanes and vessels (cruise ships and other passenger vessels) and in hybrid mode, complementing existent terrestrial connectivity in trains and other vehicles (buses, trucks or future driverless cars).

Particularly, in the case of future driverless cars, satellite role would be to provide live broadcast and multicast streams for the passengers when in remote roads, through a phased-array antenna mounted on the rooftop of the car. In that case, a low-capacity moving platform is envisaged since the computing and storage capabilities for MEC, Caching, etc. functions are expected to be smaller or more costly for a car.
For preliminary analysis, it is considered that the moving platform is an airplane with edge computing and storage capabilities and proposing media content and live streaming to passengers.

The following technical requirements associated to this SaT5G scenario 4a will be considered for the network deployment analysis:

- Connectivity in any geographical location of the aircraft;
- Minimum guaranteed satellite capacity in all weather conditions and geographical locations;
- Multi-satellite connectivity at the same time (e.g. GEO and MEO);
- Dynamic satellite capacity allocation;
- Secure connection end-to-end;
- Virtualised satellite resources advertised through an abstraction layer that exposes the satellite connection (e.g. GEO or MEO) in terms of a capacity bundle.

### 4.3.2 Negotiation scheme between the actors for SaT5G scenario 4a

Following the role model described in section 4.1, and inspired from the example of workflow represented in Figure 4-3, a negotiation process superintended by the broker entity can be initiated between different SNOs and different MNOs that serve the airline company. The negotiation model may directly include the airline or the company in charge of the airplane network management. At the end of the negotiation process, the broker shall send a report to that may include different options depending on the stakeholders involved.

One case of interest consists in the SNO that reaches an agreement directly with the airline company. This model to lease the satellite capacity is very common as of today. Different options are possible, and the report should include the following minimum set of information:

- Minimum satellite bandwidth with a fixed price per month/semester/year depending on the business model;
- Fixed amount of satellite bandwidth for a fixed price per month/semester/year depending on the business model;
- Minimum satellite bandwidth granted with a fixed price plus dynamic bandwidth allocation in a pay-as-you-go model.

Another relevant case consists in different SNOs and multiple MNOs that enter in a model agreement to lease the satellite capacity. Also in this case, different options are possible and the report issued by the broker should include the following minimum set of information:

- Multi-satellite network bandwidth availability depending on geographical location;
- MNO-X is notified with a minimum guaranteed bandwidth granted for a fixed price per month/semester/year by SNO-Y plus dynamic bandwidth allocation based on a cost model where allowed by national regulations;
- For all satellite operators, MNO-X is notified with a minimum guaranteed bandwidth provided by SNO-Y granted with a fixed price per month/semester/year for the satellites that can support Service Class-Z;
- For all satellite operators, MNO-X is notified with a fair dynamic bandwidth allocation in SNO-Y network in a pay-as-you-go model;
- Best effort model notification in case resources cannot be guaranteed.

### 4.3.3 Integrated backhaul architecture for SaT5G scenario 4a

Considering the SaT5G implementation options and the requirements associated to the SaT5G scenario 4a, Table 4-2 identifies the required high-level functionalities to be implemented in the integrated network. As stated in Section 4.1, the considered implementation options are RN based on untrusted non-3GPP access and TN based on 3GPP system specifications since the targeted term of period is currently short to medium.
Table 4-2: SaT5G scenario 4a analysis

<table>
<thead>
<tr>
<th>SaT5G High level functionality</th>
<th>Scenario requirement</th>
<th>Role model requirement</th>
<th>RN based Implementation option requirement</th>
<th>TN based Implementation option requirement</th>
</tr>
</thead>
<tbody>
<tr>
<td>Registration / Connexion</td>
<td>Yes</td>
<td>Yes (for accounting and billing)</td>
<td>Yes (since RN managed by the 3GPP NMS)</td>
<td>No</td>
</tr>
<tr>
<td>gNB/RN roaming</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Resource allocation</td>
<td>Yes (to meet the requirements of)</td>
<td>No/yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>QoS adaptation</td>
<td>Yes (to meet the requirements of)</td>
<td>No</td>
<td>Yes (mapping to satellite resources)</td>
<td>Yes (mapping to satellite CoS and then to satellite resources)</td>
</tr>
<tr>
<td>RAN aspect (incl. mobility management and beam handover)</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Multilink management</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Support network slicing</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Support of MEC</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
</tbody>
</table>

The consideration of this table can lead to a selection of the features and the architecture of the designed backhaul network.

Figure 4-6 provides the detailed backhaul architecture for SaT5G scenario 4a with RN implementation option based on untrusted non-3GPP access.

![Figure 4-6: Backhaul architecture for SaT5G scenario 4a with RN implementation option](image-url)
Figure 4-7 provides detailed backhaul architecture for SaT5G scenario 4a with TN implementation option based on untrusted non-3GPP access.

Figure 4-7: Backhaul architecture for SaT5G scenario 4a with TN implementation option
5 Recommendations for standardisation and for WP4 research pillars

5.1 Recommendations for standardisation

In the frame of SaT5G, and in particular in this document, several new functionalities are introduced in order to enable the mutual integration of the satellite and terrestrial networks. A part of these functionalities are already addressed in SDOs, currently studied by 3GPP and some need to be introduced. SaT5G is then facing three different level of standardisation:

- Already standardised;
- In the standardisation process;
- Out of scope of the standardisation.

These various states will probably lead to various impacts, in particular in 3GPP. This section introduces the potential impacts deriving from the SaT5G new functionalities.

Mobility:

The work shown in Section 3.4 has highlighted that mobility management is extremely relevant in the 5G integrated satellite-terrestrial network that will require the development of suitable mobility management procedures that can cope with the new use cases and concepts. In the Release 15 3GPP has devoted already effort with regards to mobility management and handover procedures considering the features of the 5G RAN and 5G CN. However, as 5G technology will be further developed through subsequent releases, handover mechanisms will require more studies. It is worth noticing also that the higher modularity and flexibility of the 5G system with respect to previous generations of mobile network technology through SDN and NFV, as well as the heterogeneous service types that the 5G network must support, creates a larger set of case studies for the network that has to be analysed. An exemplary case is provided by the relay node based backhaul implementation mode. This has been already introduced in 4G and its importance triggered several studies to extend the relay features also to mobile relaying scenarios. A similar path can be envisaged in the 5G technological landscapes for the IAB Donor-IAB node system applied to mobile relaying configurations. Therefore, SaT5G activities undertaken in scenarios with moving platforms leave room for possible contributions to 3GPP Release 16 and even 17.

Quality of service:

One of the key challenges for satellite integration into the future 5G networks is the convergence of the QoS treatment between both networks. This issue could be tackled with standardisation guidelines. A seamless integration could be reached by means of the same QoS model (i.e. 5G QoS model) and therefore a similar granularity. QoS treatment is also highly related to resource allocation which means that supporting 5G NR over satellite will probably enable a convergence in QoS treatment in an integrated network.

Standardisation could firstly focus on a mapping between satellite CoS and 5QI. This will be a major convergence enabler in terms of QoS treatment in a short or mid-term. Then, the next step will be a generalisation of the 5G NR and all the related impacts (e.g. resource allocation). But, this option seems to be for a long-term timescale.

In this standardisation perspective, it would be valuable to provide contribution to the relevant working group in order to push the integration of QoS adaptation. The 3GPP Working Group RAN may be appropriate to introduce the mapping through the related TS 37.324 (SDAP).

Core network protocols:

Protocols will be the point where the impact of satellite latency will be directly perceived since such latency can originate a time out. This can be an issue for mobility management as analysed in section 3.4.3.1 but also for various protocal involved either at the control or user plane, and, in general, every protocol with timers could be impacted. In a standardisation perspective, it would be relevant to study the impact of the typical timer values when applied to non-terrestrial domain and to promote satellite oriented timers if necessary. The working groups CT1 and CT4 may be the most appropriate groups for addressing this topic.
**Multilink:**

The section 3.5 has highlighted the multilink relevance. Great work has already been done in 3GPP to standardise network function in charge of traffic steering, splitting and switching (referred as AT3SF). This work has to be pushed further in order to enable the multilink support by the different backhaul implementation options. First, a RN-AT3SF would introduce an integration of the multilink by the RN. It would imply the support of relevant protocols enabling the T3S. Secondly, the multilink support by transport network based backhaul implementation would be achieved by a hybrid Multiplay Function (referred as HMF). This network function would be in charge of T3S between the terrestrial network and the satellite network. In this case, this function would embodies a role of a backhaul traffic steering, splitting and switching function (BT3SF) which has to support the relevant protocol stack (e.g. MPTCP).

In this perspective it would be interesting to provide a contribution to 3GPP; the most relevant way may be to contribute to the TR 23.793 release 16 which actually addresses access traffic steering, switching and splitting for the future 5G system.

**Generic gNB:**

SaT5G use case 4 introduces several moving platform scenarios. In this context, it would be interesting to push the concept of a generic gNB. This perspective could be addressed by 3GPP Working Group SA2 in order to derive a system architecture including the new generic gNB concept.

### 5.2 Recommendations for WP4 research pillars

In parallel of standardisation pondering, this document aims to provide some recommendations to the WP research pillars as well.

**QoS adaptation guidelines:**

This QoS adaptation is a step forward in harmonisation of Satcom with 5G control and user planes which will likely ensure better interconnection between the two systems. In this perspective, in the frame of the **WP4.4**, it would be interesting to develop the QoS adaptation mechanisms highlighted in section 3.3, and evaluate the relevance of RRC (for control plane) and SDAP (for user plane) protocols to manage such adaptation.

**BT3SF and RN-AT3SF:**

Traffic steering, splitting and switching are key enablers for the network management. In this perspective, it would be interesting to prototype HMF functionalities in order to study the system behaviour when several links are available. Multilink management through links with such various characteristics would be a relevant development in the frame the **WP4.3**.

**Support of network slicing:**

Following the definition highlighted in section 3.6, a common interface could be useful to expose which level of slicing support each part (e.g. 5GC, RAN, NTN GW, NTN terminal) can handle. In the frame of **WP4.2**, an integrated network management and orchestration design could be developed and the establishment of detailed interfaces to expose would be a major breakthrough.

**Development of high level features:**

This document focuses on the definition of high level features. In the frame of **WP4.1**, it would be valuable to implement/develop these features as VNFs in order to easily integrate them into a network management and orchestration system investigated in **WP4.2**.
6 Conclusion

The 5G satellite-terrestrial integration has been analysed throughout this document with the perspective of high level functionalities to update/develop, both in the frame of terrestrial and satellite network. The analysis has come out with the specification of the following features to support the backhaul of terrestrial network by a satellite system:

- Registration, connection and roaming management of the gNB or the relay node;
- Management of the end to end quality of service;
- Mobility management and handover in the satellite systems;
- Multilink management;
- Support of network slicing and requirement for network function virtualisation.

These functionalities integrated into the different backhaul implementation options have led to the specification of detailed backhaul architectures for each of the five identified backhaul implementation options: relay node based on untrusted non-3GPP access, relay node based on trusted non-3GPP access, relay node based on 3GPP access, transport network non-based on 3GPP system specifications and transport network based on 3GPP system specifications.

Some of these functionalities are therefore implemented by different means, considering the specifications of the backhaul implementation options and the level of integration and coordination between the satellite domain and the terrestrial domain. It has been discussed how generally speaking, for the backhaul implementation options based on the relay node, the functionalities will be typically implemented in the integrated network and will be managed by a single network management system (typically the terrestrial 3GPP NMS); while for the transport network based backhaul implementation options, a coordination between the terrestrial 3GPP NMS and the satellite NMS will be required in order to properly operate such functionalities.

Table 6-1: Requirement for the implementation of high-level features for 5G satellite and terrestrial integration

<table>
<thead>
<tr>
<th>High-level features applied to integrated satellite and terrestrial 5G network</th>
</tr>
</thead>
<tbody>
<tr>
<td>Satellite backhaul implementation options based on relay node</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>3GPP NMS</th>
<th>3GPP NMS</th>
</tr>
</thead>
<tbody>
<tr>
<td>LEO</td>
<td>LEO</td>
</tr>
<tr>
<td>HAPS</td>
<td>HAPS</td>
</tr>
<tr>
<td>GEO</td>
<td>GEO</td>
</tr>
<tr>
<td>C</td>
<td>Q/V</td>
</tr>
<tr>
<td>MEO</td>
<td>MEO</td>
</tr>
<tr>
<td>Registration, connection and roaming management</td>
<td>Registration, connection and roaming management, resource allocation and orchestration, QoS, Mobility management, multilink, slicing and virtualization</td>
</tr>
<tr>
<td>The RN reuses the registration and connection procedure as defined for the UE in 3GPP RN acts as a proxy for the UE registration and connection to the 5G core network</td>
<td>The UE registration, connection and roaming is not affected The “generic gNB” performs the registration to the 3GPP core that will serve</td>
</tr>
<tr>
<td>Resource allocation and orchestration</td>
<td>Resources manage by 3GPP NMS, including satellite resources</td>
</tr>
<tr>
<td>Management of the end to end quality of service</td>
<td>Mapping from 5QI (3GPP identifier of the quality of service) to sat resources</td>
</tr>
<tr>
<td>Mobility management and handover in the satellite systems</td>
<td>Mobility and handover, including the satellite beam handover are managed by the 3GPP NMS</td>
</tr>
<tr>
<td>Multilink management</td>
<td>The AT3SF specification for traffic steering, splitting and switching for the UE is adapted for the RN to enable multi-link-based backhaul implementation</td>
</tr>
<tr>
<td>Support of network slicing and requirement for network function virtualisation</td>
<td>E2E slicing supported and managed by 3GPP management, assuming that NTN domain management is integrated in 3GPP NMS</td>
</tr>
</tbody>
</table>

The Table 6-1 represents the identified high level functionalities with their implementation requirements for the backhaul implementation options based on RN in one hand and based on transport network in the other hand.

Considering these functionalities, a network deployment analysis has been performed in order to propose a detailed network architecture that meets the requirements of a given role model (specified in WP2.5 – Business Modelling), for a selected SaT5G scenario. The role of a broker which is responsible of trading the resources from different systems has been further clarified from a technical standpoint and a workflow for the SaT5G scenarios 2b (satellite backhaul to individual cell towers and 4a (Updating content for on-board systems and grouped media request by the moving platform company) has been proposed. It has been shown that the broker will leverage on NFV for handling an efficient network management and orchestration, by means of a federator which is able to discuss with the NMSs of various network systems, including terrestrial and satellite.

The constraint of a mid-term timescale considered in the role model and the selection of the required functionalities to address the requirements of SaT5G scenarios 2b and 4a have led to the definition of dedicated network design for each of these scenarios, based on two backhaul implementation options: relay node based on untrusted non-3GPP access and transport network based on 3GPP system specifications.

The requirement of a “generic NTN terminal” has been specified: NTN terminal shall be able to connect to any NTN system (GEO from different SNOs, MEO, HAPS, etc.), and it can be implemented either by the combination of various terminals connected to each NTN system, or by a flexible and powerful terminal being able to switch between NTN systems or even being simultaneously connected to several NTN systems.

The requirement of a “generic gNB” has also been specified for the backhaul implementation options based on transport network: the so-called “generic gNB” shall be able to connect to any 3GPP 5G Core from any MNO without requiring a specific attachment or belonging to a particular MNO. This aspect needs to be defended and pushed in the appropriate 3GPP standardisation groups in order to allow a third-party (e.g. city council, private company, airline,...) to deploy an agnostic gNB which, through a broker, can be attached at convenience to the most appropriate 5G Core. This particular
situation has been investigated for the specification of the detailed network to address the SaT5G scenario 4a: the airline company in this case has been identified as the third-party.

Finally, all the analysis performed throughout the document have led to some valuable recommendation of 5G standardisation, as for the “generic gNB” presented above, and to the definition or amendments of research pillars for the WP4 – Research to Prototype Development.

The analysis performed throughout the part A and B of document have led to some valuable recommendation for 5G standardisation contributions, and to the definition or amendments of research pillars for the WP4 – Research to Prototype Development.

The requirement for standardisation work includes:

- The adaption of UE specification for the implementation options based on RN, in order to perform: the management and registration, the mobility management of the RN, the multilink management using an adaptation of the ATSSS and the resource allocations to support end to end network slicing;
- The introduction of generic gNB concept in 3GPP in order to allow the introduction of a third-party being able to deploy an agnostic gNB which, through a broker, can be attached at convenience to the most appropriate 5G Core;
- Highlight and specify the coordination requirements between the 3GPP NMS and the TN NMS for the TN resource allocation, the TN support of network slicing and the exposure of the TN status in order to provide feedbacks to the 3GPP NMS regarding the support of a specific required QoS.

In WP4, further work will be done on the following topics:

- NFV, network slicing requirements and management an orchestration can be further investigated in WP4.1 – Implementing 5G SDN and NFV in Satellite Networks and WP4.2 – Integrated Network Management and Orchestration in line with slice management approaches described in this deliverable (and also with D3.3);
- Multilink management in WP4.3 – Multilink and heterogeneous transport: work may include development of the adaptation of ATSSS for the RN (RN-ATSSS) and analysis of the distribution protocol (MP-TCP and MP-QUIC);
- Development of the QoS and definition of the mapping table in WP4.4 – Harmonisation of Satcom with 5G Control and User Planes.
7 Bibliography


[8] iDirect, “8 Essentials to Implementing Backhaul over Satellite for Mobile Operators”.


[10] TR 103 611, Seamless integration of sat and/or HAPS systems into 5G system, ETSI, 2018.


D3.2
Integrated SaT5G Detailed Network Architecture
Part B – Caching and Multicast Architectures

<table>
<thead>
<tr>
<th>Topic</th>
<th>H2020-ICT-07-2017</th>
</tr>
</thead>
<tbody>
<tr>
<td>Project Title</td>
<td>Satellite and Terrestrial Network for 5G</td>
</tr>
<tr>
<td>Project Number</td>
<td>761413</td>
</tr>
<tr>
<td>Project Acronym</td>
<td>SaT5G</td>
</tr>
<tr>
<td>Contractual Delivery Date</td>
<td>M14 Final</td>
</tr>
<tr>
<td>Actual Delivery Date</td>
<td>17 December 2018</td>
</tr>
<tr>
<td>Contributing WP</td>
<td>WP3.3</td>
</tr>
<tr>
<td>Project Start Date</td>
<td>01/06/2017</td>
</tr>
<tr>
<td>Project Duration</td>
<td>30 months</td>
</tr>
<tr>
<td>Dissemination Level</td>
<td>Public</td>
</tr>
<tr>
<td>Editor</td>
<td>BPK</td>
</tr>
<tr>
<td>Contributors</td>
<td>ADS, AVA, BPK, BT, GLT, i2Cat, iDirect, QUO, SES, TNO, UOS</td>
</tr>
</tbody>
</table>
### Document History

<table>
<thead>
<tr>
<th>Version</th>
<th>Date</th>
<th>Modifications</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>00.01</td>
<td>2018/03/20</td>
<td>Document creation and initial delivery to WP3 contributors</td>
<td>BPK</td>
</tr>
<tr>
<td>00.99</td>
<td>2018/11/03</td>
<td>Document finalization. WP3 Internal Review: Srikant Ravuri (QUO), Simon Watts (AVA), Oriol Vidal (ADS)</td>
<td>BPK</td>
</tr>
<tr>
<td>01.00</td>
<td>2018/12/12</td>
<td>Final Version. Reviewed by Jens Krause (SES), Marlies Van der Wee (IMEC), Michaël Fitch, Georgia Poziopoulou (AVA)</td>
<td>BPK</td>
</tr>
</tbody>
</table>

### Contributors

<table>
<thead>
<tr>
<th>Name</th>
<th>Organisation</th>
<th>Contributions include</th>
</tr>
</thead>
<tbody>
<tr>
<td>Maël Boutin</td>
<td>BPK</td>
<td>Main editor</td>
</tr>
<tr>
<td>Simon Watts</td>
<td>AVA</td>
<td>Amendments to various sections, editorial changes</td>
</tr>
<tr>
<td>Christos Politis</td>
<td>SES</td>
<td>Section 5.2: Multicast SoA</td>
</tr>
<tr>
<td>Konstantinos Liolis</td>
<td>SES</td>
<td>Section 5.2: Multicast SoA</td>
</tr>
<tr>
<td>Oren Markovitz</td>
<td>GLT</td>
<td>Section 4.2: Device to device caching</td>
</tr>
<tr>
<td>Ning Wang</td>
<td>UOS</td>
<td>Section 6.2: Prefetching Workflows</td>
</tr>
<tr>
<td>Salva Diaz</td>
<td>BT</td>
<td>Section 5.3: Multicast in 5G</td>
</tr>
<tr>
<td>Michael Fitch</td>
<td>BT</td>
<td>Section 5.3: Multicast in 5G</td>
</tr>
<tr>
<td>Keith Briggs</td>
<td>BT</td>
<td>Section 5.3: Multicast in 5G</td>
</tr>
<tr>
<td>Pouria Sayyad Khodashenas</td>
<td>i2Cat</td>
<td>Section 5.5: Multicast for 5G VNF Update</td>
</tr>
<tr>
<td>Lucia D’Acunto</td>
<td>TNO</td>
<td>Section 6.2.1: Prefetching with DASH SAND</td>
</tr>
<tr>
<td>Joe Cahill</td>
<td>iDirect</td>
<td>Section 4.1: ETSI MEC</td>
</tr>
<tr>
<td>Oriol Vidal</td>
<td>ADS</td>
<td>Internal review before final version</td>
</tr>
<tr>
<td>Srikant Ravuri</td>
<td>QUO</td>
<td>Internal review before final version</td>
</tr>
</tbody>
</table>
# Table of Contents

List of Figures...................................................................................................................................... 5  
List of Tables ........................................................................................................................................... 6  
List of Acronyms..................................................................................................................................... 7 

1 Introduction ...................................................................................................................................... 8  
   1.1 Scope ...................................................................................................................................... 8  
   1.2 Relationship to other work ................................................................................................. 8  
   1.3 Document structure .......................................................................................................... 9  

2 Content Delivery: CDN architectures ............................................................................................. 10  
   2.1 Overview ........................................................................................................................... 10  
   2.2 CDN Topology .................................................................................................................. 10  
   2.3 Over the top....................................................................................................................... 11  
      2.3.1 Adaptive formats .................................................................................................... 12  

3 Caching overview: state of the art ................................................................................................. 13  
   3.1 Caching based on content popularity ............................................................................. 13  
      3.1.1 Cache tiering ............................................................................................................. 13  
      3.1.2 CDN Intelligent Edge caching ............................................................................... 13  
   3.2 Predictive caching ............................................................................................................. 14  
   3.3 Prefetching of segments of DASH/HLS video content .................................................. 15  
      3.3.1 Principle .................................................................................................................. 15  
      3.3.2 MPEG DASH SAND ............................................................................................. 15  

4 Caching implementation within 5G ................................................................................................. 17  
   4.1 ETSI MEC presentation ...................................................................................................... 17  
      4.1.1 ETSI MEC Overview ............................................................................................... 17  
      4.1.2 ETSI MEC support for Content Delivery Networks (CDN) ................................ 18  
   4.2 Device-To-Device caching ................................................................................................. 19  
      4.2.1 Device to Device communication - overview ......................................................... 19  
      4.2.2 Implementing local caching service using D2D and satellite multicast ............... 19  
      4.2.3 Advantages of D2D caching with Satellite .............................................................. 20  

5 Multicast as transport ......................................................................................................................... 21  
   5.1 Live delivery through Multicast Adaptive Bit Rate (ABR) ............................................... 21  
      5.1.1 ABR protocols .......................................................................................................... 21  
      5.1.2 Multicast ABR (mABR) ............................................................................................ 21  
   5.2 Multicast on satellite link, state of the art ........................................................................ 24  
      5.2.1 Overview .................................................................................................................. 24  
      5.2.2 IP multicast .............................................................................................................. 25  
      5.2.3 IP multicast addressing .......................................................................................... 25  
      5.2.4 Multicast group management: IGMP ..................................................................... 25  
      5.2.5 IP multicast routing ............................................................................................... 26  
      5.2.6 Reliable multicast content distribution protocol .................................................... 27
5.2.7 Reliable multicast protocols over satellites ............................................................... 28
5.2.8 Satellite multicast for linear distribution of video programs ....................................... 29
5.3 Multicast in 3GPP Networks .......................................................................................... 30
5.3.1 4G multicast implementation (eMBMS) ................................................................ 30
5.3.2 5GxCast Point to Multipoint architecture ............................................................... 35
5.3.3 Convergence of wireline IP Multicast in 5G .............................................................. 38
5.4 Multicast for content caching ......................................................................................... 39
5.5 Multicast for 5G VNF update and configuration ............................................................ 41

6 SaT5G Caching Architecture .......................................................................................... 43
6.1 Caching Workflows ..................................................................................................... 43
6.1.1 Predictive and analytics-based caching ..................................................................... 43
6.1.2 Caching orchestrated by MNO ................................................................................ 43
6.1.3 Caching orchestrated externally (e.g. CDNaaS) ......................................................... 47
6.1.4 IP Multicast for caching ........................................................................................... 50
6.2 Prefetching workflows .................................................................................................. 51
6.2.1 Enhancing prefetching workflow with MPEG DASH SAND ..................................... 53
6.3 SaT5G mABR Architecture for live delivery ................................................................. 54

7 Summary and Conclusions ............................................................................................ 57
7.1 Scope ............................................................................................................................. 57
7.2 Content distribution ..................................................................................................... 57
7.3 Content pre-positioning through caches ..................................................................... 58
7.4 Caching within ETSI ................................................................................................... 58
7.5 Multicast transport ...................................................................................................... 59
7.6 Satellite caching architectures ...................................................................................... 61
7.6.1 Caching orchestrated by the MNO ......................................................................... 61
7.6.2 Prefetching mechanisms .......................................................................................... 62
7.6.3 SaT5G mABR Architecture for live delivery ........................................................... 63

8 References ..................................................................................................................... 65
Annex A – Digital Rights Management .............................................................................. 68
List of Figures

Figure 2-1: High level content distribution architectures ................................................................. 10
Figure 2-2: Netflix OpenConnect topology [2] .................................................................................. 11
Figure 2-3: Over the top delivery ........................................................................................................ 11
Figure 3-1: Correlation between video popularity and number of views [5] ...................................... 14
Figure 4-1: ETSI MEC [11] ................................................................................................................ 18
Figure 4-2: Edge computing [11] ........................................................................................................ 18
Figure 4-3: Content update to RAN over satellite ............................................................................. 20
Figure 5-1: Possible mABR implementation ....................................................................................... 22
Figure 5-2: Multicast ABR Workflow .................................................................................................. 23
Figure 5-3: Illustrating the cost benefits of satellite multicast over terrestrial multicast distribution ... 24
Figure 5-4: Linear distribution of video programs to end user devices through satellite (Source: [34]) 30
Figure 5-5: MBMS definition [36] .................................................................................................... 31
Figure 5-6: eMBMS architecture deployment ..................................................................................... 32
Figure 5-7: U-Plane through the BM-SC ............................................................................................ 33
Figure 5-8: C-Plane through the MCE ................................................................................................ 33
Figure 5-9: Timeline for multicast services ......................................................................................... 34
Figure 5-10: 5GxCast Architecture alternative 1 [38] ....................................................................... 36
Figure 5-11: 5GxCast Architecture alternative 2 [38] ...................................................................... 37
Figure 5-12: 5GxCast Architecture alternative 3 [38] ...................................................................... 38
Figure 5-13: 5G-RG connected to 5GS through NGRAN [41] ............................................................ 39
Figure 5-14: Multicast service support [41] ....................................................................................... 39
Figure 5-15: CDN Topology ............................................................................................................... 40
Figure 5-16: NS lifecycle actions ....................................................................................................... 41
Figure 6-1: MNO Operated caching: high level architecture .............................................................. 44
Figure 6-2: MNO Operated caching: logical blocks diagram ............................................................. 44
Figure 6-3: MNO Operated caching: caching workflow ...................................................................... 46
Figure 6-4: Third party caching orchestration: high level architecture ............................................ 47
Figure 6-5: Third party caching orchestration: logical block diagram ............................................ 47
Figure 6-6: Third party caching orchestration: caching workflow ..................................................... 49
Figure 6-7: IP Multicast architecture for content caching ................................................................. 50
Figure 6-8: IP Multicast for pre-caching workflow .......................................................................... 50
Figure 6-9: Prefetching architecture and workflow (Scenario 1) ....................................................... 52
Figure 6-10: Prefetching architecture and workflow (Scenario 2) ....................................................... 53
Figure 6-11: Prefetching architecture and workflow with SAND (Scenario 1) Architecture for live content delivery ......................................................................................................................... 54
Figure 6-12: SaT5G mABR Architecture ............................................................................................ 55
Figure 6-13: mABR SaT5G Workflow ............................................................................................... 55
Figure 7-1: High level content distribution architectures ................................................................... 57
Figure 7-2: Edge computing [11] ....................................................................................................... 59
Figure 7-3: Possible mABR implementation ....................................................................................... 60
Figure 7-4: Illustrating the cost benefits of satellite multicast distribution ...................................... 60
Figure 7-5: Linear distribution of video programs to end user devices through satellite (Source: [34]) 61
Figure 7-6: MNO Operated caching: high level architecture ............................................................. 62
Figure 7-7: IP Multicast architecture for content caching as a 5G service ........................................ 62
Figure 7-8: Prefetching architecture (Scenario 1) .............................................................................. 63
Figure 7-9: Prefetching architecture and workflow (Scenario 2) ....................................................... 63
Figure 7-10: SaT5G mABR Architecture ........................................................................................... 64
Figure A-1-1: Marlin DRM operation ............................................................................................... 70
Figure A-1-2: Verimatrix full VCAS Architecture .............................................................................. 72
List of Tables

Table 1-1: Relationship between WP3.3 and other SaT5G WPs ........................................................... 8
Table 5-1: Summary table for Reliable Multicast protocols (Source: [28], [30]) ........................................ 29
# List of Acronyms

<table>
<thead>
<tr>
<th>Acronym</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>AF</td>
<td>Application Function</td>
</tr>
<tr>
<td>ABR</td>
<td>Adaptive Bitrate</td>
</tr>
<tr>
<td>API</td>
<td>Application Programming Interface</td>
</tr>
<tr>
<td>BM-SC</td>
<td>Broadcast-Multicast Service Centre</td>
</tr>
<tr>
<td>CDN</td>
<td>Content Delivery Network</td>
</tr>
<tr>
<td>CPU</td>
<td>Central Processing Unit</td>
</tr>
<tr>
<td>DASH</td>
<td>Dynamic Adaptive Streaming over HTTP</td>
</tr>
<tr>
<td>DANE</td>
<td>Dash Aware Network Element</td>
</tr>
<tr>
<td>DN</td>
<td>Data Network</td>
</tr>
<tr>
<td>DRM</td>
<td>Digital Rights Management</td>
</tr>
<tr>
<td>E-UTRAN</td>
<td>Enhanced Universal Terrestrial Radio Access</td>
</tr>
<tr>
<td>EPC</td>
<td>Evolved Packet Core</td>
</tr>
<tr>
<td>FeMBMS</td>
<td>Further enhanced Multimedia Broadcast Multicast Service</td>
</tr>
<tr>
<td>FIFO</td>
<td>First In First Out</td>
</tr>
<tr>
<td>HD</td>
<td>High Definition</td>
</tr>
<tr>
<td>ISP</td>
<td>Internet Service Provider</td>
</tr>
<tr>
<td>KPI</td>
<td>Key Performance Indicator</td>
</tr>
<tr>
<td>LFU</td>
<td>Least Frequently Used</td>
</tr>
<tr>
<td>LRU</td>
<td>Least Recently Used</td>
</tr>
<tr>
<td>LTE</td>
<td>“Long Term Evolution”</td>
</tr>
<tr>
<td>mABR</td>
<td>Multicast ABR</td>
</tr>
<tr>
<td>MAC</td>
<td>Medium Access Control</td>
</tr>
<tr>
<td>MEG</td>
<td>Multi access Edge Computing</td>
</tr>
<tr>
<td>MF-TDMA</td>
<td>Multiple Frequency Time Division Multiple Access</td>
</tr>
<tr>
<td>MBMS</td>
<td>Multimedia Multicast/Broadcast Service</td>
</tr>
<tr>
<td>MBMS-GW</td>
<td>MBMS Gateway</td>
</tr>
<tr>
<td>MBSFN</td>
<td>Multimedia Broadcast multicast service Single Frequency Network</td>
</tr>
<tr>
<td>MCCH</td>
<td>Multicast Control Channel</td>
</tr>
<tr>
<td>MCE</td>
<td>Multi-cell/multicast Coordination Entity</td>
</tr>
<tr>
<td>NFV</td>
<td>Network Function Virtualization</td>
</tr>
<tr>
<td>NOC</td>
<td>Network Operations Center</td>
</tr>
<tr>
<td>NS</td>
<td>Network Service</td>
</tr>
<tr>
<td>GSM</td>
<td>Global System for Mobile</td>
</tr>
<tr>
<td>OAM</td>
<td>Operational and Maintenance</td>
</tr>
<tr>
<td>OTT</td>
<td>Over The Top</td>
</tr>
<tr>
<td>PCF</td>
<td>Policy Control Function</td>
</tr>
<tr>
<td>PHY</td>
<td>Physical layer</td>
</tr>
<tr>
<td>POP</td>
<td>Point Of Presence</td>
</tr>
<tr>
<td>QoE</td>
<td>Quality of Experience</td>
</tr>
<tr>
<td>QoS</td>
<td>Quality of Service</td>
</tr>
<tr>
<td>RAN</td>
<td>Radio Access Network</td>
</tr>
<tr>
<td>RLC</td>
<td>Radio Link Control</td>
</tr>
<tr>
<td>ROM</td>
<td>Receive-only Mode</td>
</tr>
<tr>
<td>RRC</td>
<td>Radio Resource Control</td>
</tr>
<tr>
<td>SC-MCCH</td>
<td>Single Cell Multicast Control Channel</td>
</tr>
<tr>
<td>SC-PTM</td>
<td>Single Cell Point to Multipoint</td>
</tr>
<tr>
<td>SMF</td>
<td>Session Management Function</td>
</tr>
<tr>
<td>STB</td>
<td>Set-Top Box</td>
</tr>
<tr>
<td>UDR</td>
<td>Unified Data Repository</td>
</tr>
<tr>
<td>UMTS</td>
<td>Universal Mobile Telephone System</td>
</tr>
<tr>
<td>VNF</td>
<td>Virtualized Network Functions</td>
</tr>
<tr>
<td>VR</td>
<td>Virtual Reality</td>
</tr>
</tbody>
</table>
1 Introduction

1.1 Scope

Research on content caching in terrestrial ISP networks has existed for more than a decade. Today, caching is a broadly accepted technology. Within mobile networks, contents can be stored at either the EPC or at the RAN (Radio access network) using traditional caching policies such as LRU, LFU, FIFO and random caching. There are proposals on applying content caching in the 5G-based cellular network environment, typically at the network edge, in order to reduce content access delay for mobile users, and reduce backhaul content traffic between the mobile core and network edge.

On top of these existing caching technology proposals, satellite brings an interesting possibility to the table: its large bandwidth and its multicast capabilities. Satellites regularly use multicast for many applications such as TV distribution and private network solutions. However, multicast end to end delivery (from Network to UE) which may be used for e.g. live channel delivery, will not be studied in the scope of this project as this is part of another 5GPPP Phase 2 project: 5GxCast [1].

In this section, we will explore how we can use multicast capabilities in 5G and apply them for content caching and software update distribution.

In this document, we will focus on the delivery of OTT media assets, however the architectures described here can easily be extended to other type of contents: like web pages, file sharing, game downloads and similar. This deliverable does not look at the many complex issues surrounding the end user access to the content through tools such as digital rights management (DRM). In 2013, Avanti made an internal review of DRM for the FP7 projects BATS – some text is reproduced in Appendix One; noting that in five years the subject matter will have evolved significantly, the BATS review found that:

“Digital Rights Management has been moving from being a key part of the distribution infrastructure, to be embedded natively in the content. [...] as the centre of the home entertainment moves the content protection to be more native in the content releasing the pressure from the infrastructure and distributing it to the technologies that can be applied to the content protection and, multiple media players used for the content consumption.”

1.2 Relationship to other work

The following table summarises the relationship with other tasks in the project:

<table>
<thead>
<tr>
<th>SWP</th>
<th>Relationship</th>
</tr>
</thead>
<tbody>
<tr>
<td>WP2</td>
<td></td>
</tr>
<tr>
<td>WP2.1</td>
<td>Scenario description</td>
</tr>
<tr>
<td>WP2.2</td>
<td>Business requirements</td>
</tr>
<tr>
<td>WP2.3</td>
<td>Operational requirements</td>
</tr>
<tr>
<td>WP2.4</td>
<td>Technical requirements</td>
</tr>
<tr>
<td>WP2.5</td>
<td>Business modelling and use case description</td>
</tr>
<tr>
<td>WP3</td>
<td></td>
</tr>
<tr>
<td>WP3.1</td>
<td>Definition of the high level architecture. This architecture served as a basis for the elaboration of the WP3.3 architecture</td>
</tr>
<tr>
<td>WP3.3</td>
<td>Definition of the edge delivery architecture. This is the SWPs leading this deliverable</td>
</tr>
<tr>
<td>WP4</td>
<td></td>
</tr>
<tr>
<td>WP4.6</td>
<td>Implementation of the architectures defined in this document.</td>
</tr>
<tr>
<td>WP5</td>
<td></td>
</tr>
<tr>
<td>WP5.3</td>
<td>Demonstration of the live delivery architecture and segment prefetching</td>
</tr>
<tr>
<td>WP5.4</td>
<td>Demonstration of the Offline caching architecture</td>
</tr>
</tbody>
</table>
### 1.3 Document structure

Within this document, a discussion on how content is currently delivered is initiated and the concept of Content Delivery Network (CDN) is introduced in Chapter 2. Content Delivery Networks (CDNs) are composed of a system of servers allocated on different points of presence (POP) in an operator's network. The servers are "super-computers" defined by their capacity of storage, processing, and input/output bitrate. These servers make the video retrieved from the encoders available in the delivery networks, and replicate the video for every end user.

Then, in Chapter 3, various caching techniques are presented. These techniques are intensively used by CDNs, and rely on several factors like: content popularity, user location, session counting and dynamic prefetching or predictions.

Taking those two sections as a basis, Chapter 4 extends the caching concept to the 5G Architecture exploring how the ETSI MEC could be used in that context. Multi-access Edge Computing (MEC) offers application developers and content providers cloud-computing capabilities and an IT service environment at the edge of the mobile network. This environment is characterized by ultra-low latency and high bandwidth as well as real-time access to radio network information that can be leveraged by applications.

Afterwards, in Chapter 5, satellite benefits that can be brought on top of this are analysed and explained. Whilst the focus is on satellite multicast capabilities, a short overview of the multicast protocols is first presented. Satellite connectivity suffers from comparatively high latency but provides high bandwidth and can be used to reach nearly any point on the globe; most importantly the same packet sent can be received at many (or even all) locations that fall within the satellite beam footprint. Satellite with the addition of IP Multicast is therefore a good candidate for a caching-based solution, where the local caches can be filled offline with popular assets.

As a final step and based on the work done in WP3.1, various architectures matching different use cases are proposed in Chapter 6. At first, two architectures dedicated to offline caching of popular content are detailed, then this is complemented with a solution providing a dynamic prefetching of media segments. Finally, the specific use case of live delivery is addressed; this solution is based on the work done on mABR by the DVB group. Those solutions will be implemented within the scope of WP4.6 and demonstrated in WP5.3 (5G Cellular and Home "Plug’n Play") and WP5.4 (5G Moving Platform).
# Content Delivery: CDN architectures

Within this section, a brief overview of what is a CDN and how it works is made. A discussion is also initiated on the Over-the-top concept on which deeply relies CDN service providers.

## 2.1 Overview

Content Delivery Networks (CDNs) are composed of a system of servers allocated to different points of presence (POP) in an operator’s network. The servers are “super-computers” defined by their capacity of storage, processing, and input/output bitrate. These servers make the video retrieved from the encoders available in the delivery networks, and replicate the video for every end user.

![Figure 2-1: High level content distribution architectures](image)

A CDN is also defined by its topology (i.e., the allocation of points of presence across a certain territory) and the rules that determine the “best” server to distribute video content.

If all video content is stored in a centralized point, then all the requests will be served from this point, and the content will take long and congested routes to reach the end users. With this approach, there are multiple drawbacks. The longer content has to travel, the more significant the cost of delivery. In addition, congestion can cause delays resulting in service degradation.

On the contrary, a Content Provider could decide to replicate all video content in regional points of presence to minimize transport costs. However, this approach increases storage costs.

The goal of the CDN consists is finding the right balance between these two extremes, relying on the popularity of content. Indeed, video consumption complies with a Pareto-type law: in a standard system, 20 percent of content represents 80 percent of viewings. These figures can vary according to video services offered and users, but the principle remains: not all content has the same popularity factor with viewers. The optimization consists of locally replicating the most popular content for a specific region only and centrally storing the other content. As the popularity of content varies over time, it leads to a complex arrangement which needs to be orchestrated by the CDN.

## 2.2 CDN Topology

POPs are deployed in strategic locations closer to the end user. End users are redirected to those POPs (through HTTP redirect, BGP routing, DNS, etc.) based on their locations, profile, etc. This allows the systems to:

- Reduce the latency before the user gets its content;
- Reduce the bandwidth usage on the ISP backbone.

POP deployment is completely dynamic, meaning that a Content Provider can grow up its topology based on user demand or marketing strategy.
The following figure, taken from [2], illustrates this by showing a part of Netflix' topology (as of 2016). We can clearly see that Netflix strategy focuses on optimizing their Network for United States of America, South of America and Western Europe.

Figure 2-2: Netflix OpenConnect topology [2]

2.3 Over the top

Over the top (OTT) is different from IPTV. OTT video content is made available to subscribers directly by the providers (e.g., producers, TV channels, or aggregators) via Web portals and apps on a variety of devices, such as computers, tablets, smartphones, and connected TVs. Hence, OTT is more of a multiscreen approach. With OTT, there is no longer a need to use the operator’s set-top box to view content. However, with OTT content delivery, the operator loses the control on how its content is distributed. While operators can guarantee bandwidth dedicated to their IPTV services, OTT video content is processed in the open Internet, similarly to any other data. OTT content is transported to the ISP through private CDNs. From this point, it is not possible to distinguish the video content from the other data in transit in the operator’s network.

Figure 2-3 Over the top delivery

Video content is the most consumed network resource on Internet. In 2014, according to a study from the broadband service company Sandvine, services such as YouTube and Netflix accounted for more than 40% of an operator’s bandwidth at peak hours. Their latest report [3] states that globally nearly 59% of traffic is OTT video and Netflix alone counts for almost 15% of the global traffic, with YouTube and Amazon Prime in second and third place jointly accounting for another 15%. They also note that “It also bears mentioning that the number of mobile operators managing video traffic by offering unlimited viewing for reduced resolutions is increasing, which is also depressing the volume of video traffic worldwide”. Cisco’s own published analysis in their VNI (Visual Network Index [4]) report states “Globally, consumer IP video traffic will be 84% of consumer IP traffic in 2021, up from 79% in 2016”.

Page 11 of 72
One other point to note is their observation that gaming related traffic is on the rise and may become significant; for example: “Most games are sold via download now, and some games are huge. Call of Duty: Black Ops 3 is 101GB, and Grand Theft Auto V is 65GB. In contrast, an hour of 4K video on Netflix is about 7GB, making a Call of Duty download equivalent to watching over 14 hours of 4K video”. The move to cloud-based gaming services (where the games console is virtualized online and the resulting screen video streamed to the player) will increase this sector’s demands. The game content downloads are candidates for edge caching, however, the cloud-based gaming services are clearly not.

Operators can optimize video delivery utilizing technologies such as transparent caching. Based on the CDN principle, transparent caching enables the most popular content to be automatically stored in the delivery network closest to the end user. However, net neutrality restrictions do require that all data on the Internet be treated equally preventing this approach to be a viable and satisfactory solution. To some degree this is not fully followed by the MNOs, for example it is possible to buy consumer packages that include streaming from selected OTT providers that is excluded from the monthly data limits.

### 2.3.1 Adaptive formats

Using an adaptive delivery format, a Content provider can decrease the network load while improving the user’s quality of experience at the same time. Thanks to this technology, when the network has been saturated to a point where it cannot transport an additional signal, the receiver requests a lower bit rate from the server, reducing the amount of data so that the signal can be delivered. Of course, the quality of the video content resulting from this request is slightly degraded, but this is more acceptable than a break in service. When the network conditions become better, the bitrate and therefore the quality can increase again accordingly.

Currently, the most used adaptive formats are Apple HLS (HTTP Live Streaming) and Microsoft HSS (HTTP Smooth Streaming). An open standard called MPEG-DASH (Dynamic Adaptive Streaming Over HTTP) exists, and is also being deployed.
3 Caching overview: state of the art

Content caching is a way to optimize content delivery. This is typically done by storing assets in various network places, closer to the end user thus optimizing both QoS and network resources. On-net caching avoids having to access content over the Internet and currently more than 50% of content delivered by BT’s network in the UK is from on-net caches.

Caching can be triggered from various criteria: session counting, localized contents, content popularity or any service specific trigger.

We can divide caching in three main algorithms:

- Generic caching based on session counting;
- Predictive caching (prefetching of content before it is requested);
- Prefetching of segments based on current sessions.

In this chapter, several techniques to cache content are presented to establish the catalog of possible solutions for edge delivery solutions addressed in WP3.2.

3.1 Caching based on content popularity

This is the most common implementation of caching. When content is requested from a streaming server, this content is automatically cached locally when a predefined threshold is reached so it can be served directly from a local storage in the future. This threshold is based on the number of session requests for this content. This drastically increases the QoS (and QoE) and optimizes the network resources.

This however has some drawbacks:

- If the threshold is not correctly defined, then it may have the side effect of decreasing the storage lifetime in the edge due to the huge number of write cycles implied.
- It increases the storage quantity on the streamers and hence increases the cost of the edge servers.

To deal with these issues, some additional intelligence may be implemented:

3.1.1 Cache tiering

To mitigate the cost increase related to storage management, some servers use a “cache tiering” feature. Basically, servers have different types of storage from the lower performance and cheapest drives to the high performing/ most expensive storage media. A streamer's internal threshold is configured to determine how many concurrent sessions are required to move content from one media to another.

Typically, three different storage media can be managed:

- SAS/SATA hard drives (can be internal, DAS or SAN) (less popular content);
- First cache level: SSD drives (medium popular content);
- Second cache level: RAM (most popular content).

3.1.2 CDN Intelligent Edge caching

CDNs implement advanced algorithms such as Intelligent Edge Caching. With this feature, it becomes possible to state that long-tail contents (content that were added to catalog long ago and which are not popular anymore) are streamed from a central location and that only popular content is cached in edge servers and streamed from there.

Streamers are elected by the CDN based on the configured topology:

- **Central server**: if the caching threshold (based on the criteria presented before) has not been reached yet, the central server is elected to deliver the session;
- **Edge servers and cache affinity:** Edge servers will be elected once the threshold has been reached, furthermore if a server has already been elected for a content then future sessions on this content will be redirected to this server in priority.

This way of distributing the sessions circumvents the two issues described in previous Section 3.1.

The trickiest part is the determination of the threshold used to decide when an asset should be streamed from the edge. To help on this, some studies ([5], [6]) found that content delivery follows Pareto’s law: in a standard system, 20% of content represents 80% of viewings. These figures can vary according to video services offered and users, but the principle remains: not all content has the same popularity factor with viewers. Such trends have been noticed in networks ranging from cellular, to user-generated content, to IPTV and VoD. As an illustration, for both YouTube and Daum (service in Korea), 10% of the most popular videos account for nearly 80% of views, while the remaining 90% account for total 20% of views as depicted in the figure below [5]:

![Figure 3-1: Correlation between video popularity and number of views [5]](image)

Content popularity tends to follow a heavy tail distribution which means that the majority of requests occur for a relatively small fraction of the content. Observations of this characteristic have been noted in data logs for various CDNs.

This can be quantified using Zipf’s law. Specifically, if we ordered the files from most to least popular at a given point in time, then the relationship governing the frequency at which the file of rank i will appear is given as follows [6]:

\[
 f(i) \propto \left( \frac{1}{i} \right)^\alpha
\]

That is, the probability of a request occurring for file \( i \) is inversely proportional to its rank, with a shaping parameter \( \alpha \). A larger \( \alpha \) implies that more requests occur for a smaller fraction of the content, making the network more amenable to a caching solution. Values of \( \alpha \) have been cited between 0.5 and 0.8 [6].

### 3.2 Predictive caching

Predictive caching is another extensive research area where, based on statistical models, the aim is to predict what/how/when/where a user (or a panel of users) will ‘play’ in order to cache (or a part of) it before the user hits play [7]. As the content is cached beforehand, the quality of the cached content can be higher since it is transferred in the background. Such an approach is typically useful for content like TV series or episodes. For example, we can assume that if a user has been watching a particular series it is very likely that this user will play the next unwatched episode. In this case the content of the next episode can be proactively cached before the user starts to consume it. In the context of SaT5G, in case a piece of content is expected to be popular at multiple network locations, the leverage on satellite backhaul links to multicast the content to the local caches to be proactively cached at those edge locations is a promising solution.

Complex algorithms can be envisioned based on several variables like (not exhaustive):
- User location;
- User profile (kid, senior...);
- Viewing history of a group of users and others.

When content becomes "popular" then it may be pushed to/pulled by the local streamers depending on the implementation.

The following paper [8] provides four implementations of predictive caching using various metrics:

- Weighted popularity of an asset in various day timeframes.
- Size of the asset. Large assets will cost more if retrieved from the origin, it is therefore desirable to have them stored in the cache
- Network topology

3.3 Prefetching of segments of DASH/HLS video content

3.3.1 Principle

In addition to the offline multicasting of popular content to the mobile edge for local access, we also investigate how satellite links can be directly involved in the real-time delivery of video content in 5G environments. It is apparent that the long propagation delay of satellite links is a significant issue that can affect the end-to-end performance of DASH (Dynamic Adaptive Streaming over HTTP) / HLS (HTTP Live Streaming) video streaming based on TCP. TCP throughput performance can suffer substantially across any end-to-end path containing heterogeneous path segments with substantially different bandwidth-delay products (BDP). Satellite networks typically include performance enhancing proxies that minimise the impact at the TCP layer unless the TCP headers are encrypted (e.g. [9]). Modern TCP stacks typically perform much better than older stacks over a GEO satellite link with ~600ms round trip delays.

In a mobile environment, the radio access network (RAN) part typically has a lower BDP value compared to the backhaul/Internet part with long delay (up to 350ms for terrestrial networks) and a fat data pipe. It can be easily inferred that the additional delay introduced by satellite links makes the situation more challenging, and indeed may induce poor end-to-end performance in DASH video delivery.

In order to reduce the impact of latency, the key idea is to leverage a multi-access edge computing (MEC) server that is responsible for intelligently pre-fetching on-demand an optimal number of DASH video segments from the original content source and make them locally available to the consumers. Specifically, the scheme leverages on the “fat” pipe of the backhaul and Internet path to establish additional TCP connections in a controlled manner to pre-fetch future segments ahead of user requests during a real-time video session. Such a solution does not require the pre-caching of video content at the mobile edge a priori, which makes it an ideal solution even for non-popular content objects. According to the experiment study [10] based on the terrestrial networks, the proposed approach is able to achieve QoE-assured 4K video delivery across the entire Internet provided that there are adequate radio resources on the RAN side.

3.3.2 MPEG DASH SAND

MPEG DASH SAND (Server and Network-Assisted DASH) is an MPEG/ISO standard (ISO/IEC 23009-5:2017) which has been developed to enable consistent and higher quality service for adaptive video streaming applications using DASH delivery. To achieve this goal, SAND establishes an asynchronous communication channel between the DASH client and a DASH-Aware Network Element (DANE), through which the DASH client obtains assistance from the streaming server and/or the network. One of the use cases of MPEG DASH SAND is “Intelligent Edge Caches for Mobile Users”, where the DANE is implemented as a MEC server enabled to cache DASH segments that DASH clients are expected to request. DASH clients can inform the MEC DANE about the next segments to be requested by sending the SAND status message “AnticipatedRequests”. This way, the MEC DANE may decide to pre-fetch the content for faster delivery. The client may also provide a deadline for specific segments, such that the MEC DANE can schedule and pace the prefetching of the segments.
Within the context of a dual 5G backhaul network connection (i.e. via satellite), the MEC DANE may also choose to pre-fetch some segments via the terrestrial link (i.e. if they are needed soon by the requesting DASH client), and the bulk of other segments via the higher bandwidth satellite link (e.g. when their due time is far enough). In this way, the load on the backhaul networks can be balanced while at the same time providing high quality streams to the DASH clients.
4 Caching implementation within 5G

In this chapter a brief overview of the existing 5G caching technologies will be presented. At first, a presentation of ETSI MEC is provided and an explanation on how this technology can leverage caching is proposed. Then D2D (Device-to-Device) caching is presented.

4.1 ETSI MEC presentation

4.1.1 ETSI MEC Overview

Edge computing as an evolution of cloud computing which brings application hosting from centralized data centres down to the network edge, closer to consumers and the data generated by applications. Edge computing is acknowledged as one of the key pillars for meeting the demanding Key Performance Indicators (KPIs) of 5G, especially as far as low latency and bandwidth efficiency are concerned.

However, not only is edge computing in telecommunications networks a technical enabler for the demanding KPIs, it also plays an essential role in the transformation of the telecommunications business, where telecommunications networks are turning into versatile service platforms for industry and other specific customer segments. This transformation is supported by edge computing, as it opens the network edge for applications and services, including those from third parties.

Multi-access Edge Computing (MEC) offers application developers and content providers cloud-computing capabilities and an IT service environment at the edge of the mobile network. This environment is characterized by ultra-low latency and high bandwidth as well as real-time access to radio network information that can be leveraged by applications.

MEC provides a new ecosystem and value chain. Operators can create new revenue streams by exposing their Radio Access Network (RAN) edge to authorized third-parties for application hosting and providing RAN-based services, allowing them to flexibly and rapidly deploy innovative applications and services towards mobile subscribers, enterprises and vertical segments.

The MEC initiative [11] is an Industry Specification Group (ISG) within ETSI. The purpose of the ISG is to create a standardized, open environment which will allow the efficient and seamless integration of applications from vendors, service providers, and third-parties across multi-vendor Multi-access Edge Computing platforms.

The ISG aims to benefit all entities within the value chain, including mobile operators, application developers, Over the Top (OTT) players, Independent Software Vendors (ISVs), telecom equipment vendors, IT platform vendors, system integrators, and technology providers. All of these parties are interested in delivering services based on Multi-access Edge Computing concepts.

The work of ETSI MEC aims to unite the telco and IT-cloud worlds, providing IT and cloud-computing capabilities within the RAN (Radio Access Network). The ISG MEC will specify the elements that are required to enable applications to be hosted in a multi-vendor multi-access-edge computing environment. MEC will enable applications and services to be hosted ‘on top’ of the mobile network elements, i.e. above the network layer. These applications and services can benefit from being in close proximity to the customer and from receiving local radio-network contextual information.
5G networks based on the 3GPP 5G specifications are a key future target environment for MEC deployments. The 5G system specifications and its Service Based Architecture (SBA) leverage the service-based interactions between different network functions, aligning system operations with the network virtualization and Software Defined Networking paradigms. These same characteristics are shared by MEC specifications. In addition, 3GPP 5G system specifications define the enablers for edge computing, allowing a MEC system and a 5G system to collaboratively interact in traffic routing and policy control related operations. MEC features together with these complementary technical enablers of the 5G system allow integration of these systems to create a powerful environment for edge computing.

In the frame of SatCom, deploying applications and delivering services at the edge of the satellite network can bring even more benefit to the satellite end user as the impact on the delay, and subsequently the user’s QoE, brings greater benefit.

### 4.1.2 ETSI MEC support for Content Delivery Networks (CDN)

ETSI MEC defines a framework for hosting and delivering services at the network edge. The ETSI MEC framework provides a secure and trusted platform at the network edge, within the trusted domain of the telecom’s operator, to host and provide services to the end user and provide additional services that are unique to the network edge (e.g. real time location, congestion information).
Local content caching and delivery, using Content Delivery Networks (CDNs) or other means, play a key role in many of the use cases identified by ETSI MEC, including:

- Video analytics;
- Location services;
- Internet-of-Things (IoT);
- Augmented reality;
- Optimized local content distribution; and
- Data caching.

The benefits of having a local CDN at the MEC location are:

- Reduced latency for accessing content;
- Backhaul savings when delivering content locally;
- Improved QoE for the end user;
- Real-time awareness for improved content delivery.

The content cache application can store the content that has been identified as frequently used or otherwise beneficial from the service point of view. Content caching applications can use information obtained from other applications to identify the content that could be cached.

Once the content cache application receives a request for content that is stored in its local cache, the application starts directing the requested content to the user equipment, which requested the content. This results in savings in the backhaul capacity as well as improvement in QoE as content can be transferred without the additional delays caused by the transport (mobile and/or satellite) link, core network and public internet. The satellite link is particularly well suited for this caching task due to its high bandwidth and its availability all around the globe.

### 4.2 Device-To-Device caching

Device-to-Device communication was introduced in LTE REL12 and enables devices located within the same cell to directly communicate over allocated spectrum. The standardization of D2D is not complete yet; however, it may offer a very simple and standard option to deploy caching services and it turn out to be a particularly well-suited technique to be married with satellite content update capabilities.

#### 4.2.1 Device to Device communication - overview

The following is a simplified overview of Device-to-device communications:

- “Serving Devices” can register their service and publish the service they offer using multicast;
- “Client Devices” that wish to use the service ask the cellular base station to allocate a slice of spectrum (time and frequency slot);
- The client device uses the allocated bandwidth to directly communicate with the serving device without utilizing the cellular infrastructure.

#### 4.2.2 Implementing local caching service using D2D and satellite multicast

This consists of three main phases:

**Content update:**

- The cache provides statistical usage telemetries to the cache manager using the cellular infrastructure
- The cache manager updates the different caches using multicast groups without overloading the cellular infrastructure

**Service advertisement:**

The cache publishes its service using D2D service advertisement to devices located in the same cell and that are in proximity to the cache device.

**Serving content:**
Devices can establish direct connection with the cache using D2D communication and load content such as multimedia. When the requested content is not stored in the cache, the client device can either fetch it directly or the cache can serve as proxy device and fetch the content. When fetching online content, the cache device uses the cellular infrastructure.

**Figure 4-3: Content update to RAN over satellite**

### 4.2.3 Advantages of D2D caching with Satellite

The combination of D2D with Satellite for content updates has the following advantages:

- Service can be deployed without overloading the cellular infrastructure due to content updates;
- Service advertisement is done using a standard 3GPP interface;
- Service and sessions between client and server can be established in a standard way without the need to implement local breakout between the base station and cache.

In the frame of Sat5G, this particular technique will not be further investigated, due to the fact that it is not yet completely defined in 3GPP. SaT5G will focus on the more traditional caching technologies where the content is cached on a content source and not on another device.
5 Multicast as transport

5.1 Live delivery through Multicast Adaptive Bit Rate (ABR)

5.1.1 ABR protocols

Nowadays, OTT content delivery is dominated by the ABR protocols. Those L4 protocols take advantage of the HTTP universality:

- Ubiquitous: implemented everywhere: browsers, smartphones, tablets;
- Ease of player development;
- Easy firewall traversal;
- Relies on TCP congestion algorithms.

The main concept of the ABR protocol is to adapt the delivery bitrate based on the user end network connectivity in real time (even when roaming from a different access network).

In unicast, ABR is done by the UE Player requesting the relevant bitrate piece/segment/part using HTTP GET to the Content Provider. When the throughput to that UE drops, the player will request lower bitrate content piece.

A variety of ABR implementations exists, among those, the most used are:

- MPEG DASH;
- HLS;
- SmoothStreaming.

All of them rely on a “top” manifest file which contains the various segments and their bitrate. The player firstly retrieves this manifest and then chooses the bitrate of the next segments to fetch from the source.

The main benefits of the ABR technology are the following:

- Works over HTTP protocol suite. This means that the packets have no difficulties going through Firewall / NAT;
- No “intelligence” required on the streamer (stateless sessions) which allows the development of high-performance streamers (over 100GB/s). This greatly enhances the overall scalability;
- The client application manages the packets which it wants to receive based on network conditions, user preferences, available computing resources etc.

5.1.2 Multicast ABR (mABR)

While ABR currently dominates the world of Over the Top (OTT) content delivery, it suffers from one serious drawback. The protocol allows for the current best scalability of static content delivery; however, the relationship between the client application and the streamer is a one-to-one. This means that if e.g. 1,000,000 users request the same live channel at the same time, 1,000,000 connections to the streamers will be made. This is where mABR comes into play.

In the case of mABR (Multicast Adaptive Bit Rate), live channel is multicast within the backbone and converted back to unicast closer to the end user device. This mABR technology is only viable for live content. VoD as per nature are requested independently by each user and therefore cannot be transported on Multicast due to a lack of Multicast resources (and their price!)
Figure 5-1: Possible mABR implementation

As shown in Figure 5-1, a single multicast connection is managed between multicast server (represented by Transcaster server in this figure) and the Multicast Gateway function within the Home Network.

This figure is a possible implementation of mABR. mABR is being standardized by the DVB Group [12]. The implementation proposed here does not reflect all the possibilities offered by the DVB standard. As an example, the conversion from multicast to unicast could be done at other points in the end-to-end chain:

- In the End User device if multicast is supported all along the path (e.g. eMBMS client);
- At the Edge of the Network (e.g. as a MEC User Plane function).

5.1.2.1 mABR generic workflow

The following figure (Figure 5-2) illustrates a generic multicast ABR workflow

In the below sequence diagram, the Multicast Gateway (Function Y) acts as a reverse proxy forwarding received requests to the CDN if necessary or serving them from its local cache. While this option is the simplest to implement, other solutions are possible:

- HTTP Redirects: The video player contacts the CDN which in turn redirects it to the Multicast Gateway. This type of solution is more Content Service Provider friendly: the CDN continues to receive the unicast Session initiation request, thus not disrupting its internal operations (redirections, authentication & authorization, analytics…). Deliverable D5.2 [32] focuses on this solution and further details the associated call flow.
- Middleware: Application Middleware has been informed (through static configuration or dynamically via an API) of the fact that Multicast Gateway should be used for this channel. This way, the middleware will ask for the packets to the Multicast Gateway instead of the Origin server

The workflow presented below is separated in two distinct unsynchronized parts: Multicast configuration and Session Workflow.
5.1.2.2 Configuration

**Note:** Configuration is updated dynamically and is not synced with the Session workflow

1. Get the popular lives: Multicast Server (function X) retrieves periodically the list of popular lives. Based on this information, function X starts multicasting the live in the NSP Network (purple arrow). Two options are here offered to the Multicast Server: stream all channel layers (one multicast per layer) or stream only a subset of the layers. The popular lives can be determined:
   a. Statically thanks to an API (the channel streaming the football match may have been set to popular in advance)
   b. Dynamically thanks to an Analytics server which would count the number of same Sessions per region, user profile etc.

5.1.2.3 Session workflow

The two flows START Session and STOP Session serve as containers for any unicast CDN specific operations. There are many possible operations: analytics, authentication and authorization etc.

1. The UE starts a streaming Session and initiates a connection to the Multicast Gateway (Function Y) to retrieve the manifest file. Depending on the implementation, the way how the video player retrieves the manifest differs
   a. Reverse proxy: the video player contacts the Multicast Gateway which forwards the request to the CDN. This is the case described in the message flow.
   b. HTTP Redirects: the video player contacts the CDN which redirects to the Multicast Gateway.
   c. Middleware: The middleware contacts directly the Multicast Gateway for this specific channel.
2. The Multicast Gateway forwards the request to the CDN and answers to the UE
3. As any regular Session the UE then asks for the layer manifest
4. The Multicast Gateway asks the CDN for this manifest and returns it back to the UE
5. UE asks for the first chunks

---

![Multicast ABR Workflow](image-url)
6. The Multicast Gateway retrieves the chunks from the CDN then joins the multicast on this live if necessary (Multicast Gateway may have already joined this multicast).

7. The Multicast Gateway caches the chunks from the multicast and forwards them to the UE on request.

It is worth noting that first steps are handled using a unicast transport. We use unicast at the initial stage in order to improve QoE when switching between channels. Indeed, joining a multicast can take up to a few seconds.

In case of a multiscreen scenario (TV, smartphone, tablet …) the unicast steps may be skipped if the chunks are already cached in the Multicast Gateway.

5.2 Multicast on satellite link, state of the art

5.2.1 Overview

Nowadays, there is an increasing demand for efficient distribution of personalized content through Internet-based networks [13]. This has led to the deployment of satellite platforms delivering high throughputs (HTS systems) or constellations of communication satellites such as the O3b system [14], which are able to satisfy the user requirements by enhancing the delivery of multimedia content. Multicasting is one of the cornerstones for the effective dissemination and distribution of personalized multimedia content in broadband networks. Applications such as audio/video streaming, online gaming, file distribution, and file downloading are based on multicast transmissions.

The cost benefits of using satellite to distribute multicast content as the number of receiving nodes increase are illustrated in Figure 5-3 below. The real world values illustrated in Figure 5-3 will be different but this simply shows that the bandwidth cost for satellite multicast remains fundamentally the same regardless of the number of nodes whereas the terrestrial multicast bandwidth costs are roughly proportional to the number of nodes. Figure 5-3 illustrates Use case 3 described in [15].

![Figure 5-3: Illustrating the cost benefits of satellite multicast over terrestrial multicast distribution](image)

Satellites have for many years regularly used multicast for many applications such as TV distribution and private network solutions. Several initiatives, driven by private companies and by consortiums, are leading to the elaboration of Multicast Adaptive Bit Rate standards (mABR) [16], [17]. These standards are not directly applicable to satellite and mobile distribution so far, since they are based on the existence of a home network equipment converting multicast back into unicast, an equipment that is not currently present in mobility. They are however being investigated in the scope of other projects (e.g. 5GxCast [1]). This function is necessary since most mobile devices are not capable of receiving a multicast feed, as it requires a new user device chipset design (as defined in LTE Broadcast 3GPP standard TS 23.246 [18]).

Further details on the State-of-the-Art in the Multicast for Satellite Networks are provided hereinafter.
5.2.2 IP multicast

IP Multicast is a networking technique, which allows a source to send data to multiple destinations simultaneously using a single transmit operation. Multicasting makes efficient use of network bandwidth over unicast transmission. In this context, satellite is the ideal technology for multicasting, given the efficiency it provides with point-to-multipoint communication, and given its global coverage, making it the more cost effective medium when there are multiple requests for the same content. The introduction of increasingly wideband IP-based satellite networks allowed the implementation of vendor specific IP multicast systems [19], which require a server at the satellite gateway and a client software at the satellite terminal to capture the multicast content. 3GPP has defined a reference architecture for Multimedia Broadcast and Multicast Services (MBMS) in TS 23.246 [18] but this technology is chosen for deployment within terrestrial networks instead of operation over satellite networks.

5.2.3 IP multicast addressing

Each terminal or host in the Internet is uniquely identified by its IP address. The first main version of IP, Internet Protocol Version 4 (IPv4), is the dominant protocol of the Internet and its successor is the Internet Protocol Version 6 (IPv6). IPv4 consists of a 32-bit address, divided into a network number and a host number, which respectively identify a network and the terminal attached to the network. The header in a normal unicast IP packet includes a source address and a destination address. Then, the destination address is used by routers to route the packet from the source to destination. This process cannot be used for multicast scenarios, because the source terminal may not know when, where and which terminals will try to receive the packet.

To overcome this issue, a range of addresses is defined for multicast purposes, where this range of addresses is only associated with a multicast group. Then, routers use these addresses to route IP multicast packets to registered users, in the multicast group. For IPv4 cases, a terminal registers in a group using the Internet Group Management Protocol (IGMP) [20], while for IPv6 cases, the Multicast Listener Discovery Protocol (MLD) [20] is adopted.

IP Multicast provides the packet routing from a central server to the receivers – a multiplicity of applications using proprietary processes and nowadays some standards-based components such as NORM [21]. Typical applications offer human and machine interfaces to move the content to the desired locations. A full description of these applications is out of scope of this deliverable which focusses on the network and transport layers.

5.2.4 Multicast group management: IGMP

The IGMP allows hosts or terminals to declare an interest in receiving a multicast transmission. There are three main types of IGMP messages:

- Membership Queries;
- Membership Reports; and
- Membership Leaves.

When a terminal wants to receive a multicast transmission, it sends an IGMP Report, which is received by the nearest satellite router. This Report specifies the IP multicast address of the group being joined. Then, the router uses a multicast routing protocol to determine a path to the source. Furthermore, the router sends occasionally IGMP Queries to terminals in its network in order to learn the state of terminals receiving multicast. When a terminal receives such a query, it sets a separate timer for each of its group memberships. When each timer expires, the terminal sends an IGMP Report to confirm that it still wants to receive the multicast transmission. Finally, when a terminal desires to finish receiving the multicast transmission it sends an IGMP Leave request. If all the members of a group in a subnet have left, the router does not forward any more multicast packets to that subnet.

\[1\] Most modern satellite terminals incorporate both modem and router functions and can be referred to as satellite routers in some documents.
IGMP behavior in satellite environments

There are three versions of IGMP, and today the satellite multicast systems are based on them. As an illustration, some COTS available products include but not limited to the following:

- SkyEdge II System [22] from Gilat;
- X7 Satellite Routers [23] from iDirect; and

In a satellite environment, multicast group management together with the transport protocols provides an efficient solution to support IP multicast. IGMP and MLD over satellite present a unique and challenging scenario, due to the large number of receivers that can be present in the satellite geographical coverage area, which could cause significant flooding of the satellite network with IGMP or MLD traffic. Therefore, a number of adaptations of the IGMP or MLD traffic should be obtained. The remaining text [25] delves into the forwarding multicast traffic that is received on the satellite interface. It should be noted that many satellite terminals do support IGMP and integrate these requests into their control fabric as multicast is seen as a key strength of satellite communications.

The Network Control Center (NCC) is responsible for all multicast transmission on the forward link. Prior to multicast data transmission on the forward link, the NCC should first configure the Feeder and the Gateway Router with entries for those multicast streams that are to be forwarded. Once configured, the Gateway Router joins the requested group at its upstream interface. The feeder encapsulates the multicast traffic and forwards it over the satellite air interface. Static and dynamic forwarding are distinguished in this case. In static forwarding, this process is completed before the satellite terminal needs to receive the multicast group. In dynamic forwarding, this process is completed the first time a satellite terminal requests to join a specific group. A satellite terminal that wishes to receive multicast traffic with a specified IP address on the forward link should first construct a layer 2 filter containing the Generic Stream Encapsulation (GSE) labels with which the multicast traffic is sent. This table may be directly mapped using the information in the Multicast mapping Table 2nd generation (MMT2) to identify the GSE address used to carry a multicast flow. Once configured, the filters forward all traffic with the label to the IP layer where the satellite terminal filters the traffic based on the IP network layer address using the information in the Multicast Forwarding Information base (MFIB).

Static forwarding over the satellite has advantages in terms of simplicity of design of the Gateway and control of the Quality of Service (QoS) offered to each multicast group. However, the approach relies on the operator determining what content is to be received at any time. While this is appropriate for pre-scheduled transmissions (such as file updates, IPTV broadcast, etc.), it is not appropriate for applications that are user-driven (such as video-on-demand, multi-party conference, collaborative working applications and service discovery). User-driven applications often cannot predetermine the set of multicast groups that will be used, and it is often not feasible to forward all multicast traffic over a satellite, irrespective of whether there are any active receivers for the given groups. Dynamic forwarding on the satellite interface is required in these cases to control the set of groups that are forwarded from the Gateway to the receivers. However, it increases the complexity of the multicast service as it necessitates more control functions to realize an effective operational service. When considering the population of MEC-based caches, a static set of multicast groups can be readily envisaged within which the cache management processes deliver content as requested, the satellite terminals using IGMP to pass the content to LAN when the cache has requested the specific content.

In architectures that have no router on the downlink side a proxy can be used to forward multicast traffic to group members.

5.2.5 IP multicast routing

Multicast routing protocols address the issue of identifying a route for data to be transmitted across a network from a source to all its destinations, while minimizing the total network resources required for this. In IP multicast, the routing table is effectively built from destinations to the sources rather than from sources to destinations, since only the source address in the IP datagram corresponds to a single physical location. A number of multicast routing protocols have been developed by the Internet Engineering Task Forces (IETF) and are described in [26].
5.2.6 Reliable multicast content distribution protocol

Multicast can be either best effort or reliable. “Best effort” means that there is no guarantee that the data sent by any multicast source is received by all or any receivers, and is usually implemented by a source transmitting User Datagram Protocol (UDP) packets on a multicast address. On the other hand, “reliable” means that the implemented algorithms ensure that all receivers of a multicast transmission receive all the sent data: this requires a reliable multicast protocol.

Within reliable multicast, there are many protocols both for terrestrial and satellite networks. However, the IETF Reliable Multicast Transport (RMT) working group recognizes two families of protocols, the Negative ACKnowledgement (NACK) based protocols and Forward Error Correction (FEC) based protocols [27].

NACK-based multicast

In NACK-based multicast, receivers should only provide negative acknowledgments informing the multicast source of reception errors or losses. The motivation for using negative acknowledgments is to keep the amount of feedback from receivers as low as possible. The less feedback a protocol requires, the more scalable it can be. The problem arises when a large number of receivers return an ACK packet for every packet they receive correctly, causing serious network congestion. Some reliable multicast protocols based on the NACK method are [28]:

- Multicast Transport Protocol (MTP/MTP-2);
- Multicast Dissemination Protocol (MDP/MDPv2); and

A critical part of these protocols is the NACK repair process. This includes the receiver’s role in detecting and requesting repair needs, and the sender’s response to such requests. The receivers mainly send unicast packet to request for retransmissions, which later multicast back to requesting receivers.

Upon reception of a repair request from a receiver in the group, the sender will initiate a repair response procedure. The sender may wish to delay transmission of repair content until it has had sufficient time to accumulate potentially multiple NACKs from the receiver set. Immediately after the sender NACK aggregation period, the sender will begin transmitting repair content determined from the aggregate NACK state and continue with any new transmission. In general, the timing of retransmissions should be dependent upon the greatest round-trip timing (GRTT), where the sender GRTT is an estimate of the worst-case round-trip timing from a given sender to any receivers in the group. More details can be found in [29].

FEC-based multicast

In FEC-based multicast, receivers can either: receive two separate streams, one with application data and another with parity data or receive a single stream of encoded symbols, in order to avoid depending on specific data packets but rather on the number of different packets received. FEC coding is a well-known technique for protecting data against corruption. Hence, one immediate benefit of packet level FEC coding is the reduction in the number of lost packets, minimizing with way the number of feedback reports as well as the need for retransmissions.

Hybrid NACK/FEC-based multicast

Besides these two families, there is also a number of hybrid reliable multicast protocols that combines both approaches. These hybrid protocols are supposed to incorporate the best from both families. Some reliable multicast protocols based on the hybrid NACK-FEC method are [28]:

- Multicast File Transfer Protocol (MFTP);
- Reliable Multicast data Distribution Protocol (RMDP); and
- Restricted Reliable Multicast Protocol (RRMP).
It can be noted that some proprietary systems offer the choice of methods and may allow different techniques to be combined; for example, BE2 + 15% FEC overhead.

**Best effort transmissions**

A common practice for larger satellite content delivery networks is to simply send the content multiple times (known as best effort or BE); for example, if the content is sent out twice this can be referred to as BE2. A statistical analysis would be performed based on typical content size, the link packet error rates and the number of devices to determine the amount of content that would need to be requested, if this approaches the size of a full retransmission then BE2 becomes sensible and significantly reduces the amount of ACK/NACK traffic on the return path.

This technique can be combined very effectively with FEC-based multicast.

This is not strictly reliable in the pure sense but nevertheless has proved to be very effective in commercial deployments – the use of this however is covered by NDAs so specific references can’t be provided.

**Fountain codes**

Fountain codes (also known as rateless erasure codes) - The transmitter using Fountain codes basically adds redundant information to a set of k blocks generating a stream that is slightly larger than the original stream. The receiver is able to recover the original stream of blocks from any subset of the received blocks that is only slightly larger than the number of source blocks.

Raptor codes are the most common standard implementation of Fountain code. It has been adapted and implemented in 3GPP MBMS for broadcast file delivery and streaming services, and in DVB-H IPTV.

### 5.2.7 Reliable multicast protocols over satellites

As mentioned earlier, to provide reliability, a protocol needs to identify the packet which failed to reach a given destination and this can be obtained with NACK-based multicast. However, for a satellite multicast application, due to high error rates of the satellite links and large number of receiver-sets, even NACK-based feedback may lead to an implosion problem. Moreover, feedback packets need to be transmitted through the shared satellite uplink, causing waste of already scarce satellite resources. Because of these issues, the multicast protocols couple their feedback scheme with supporting mechanisms such as feedback suppression.

Feedback suppression is applicable if NACK-based feedback reports are multicast to the network. This allows receivers to listen to the feedback reports and to suppress their own feedback reports if the received NACK report is for the same missing packet. One known protocol which combines NACK-based repair strategy and feedback suppression is the NACK-Oriented Reliable Multicast (NORM) transport protocol [30], [29].

**NORM transport protocol**

The NORM protocol is designed to provide reliable, efficient, scalable, and robust transport of large amounts of data over an IP multicast network. The protocol uses:

- a NACK-based repair strategy, where the receivers request repair of missing data by sending non-acknowledgments when they discover packet loss,
- receiver feedback suppression mechanisms, that allow the receivers to cancel superfluous feedback messages to the sender, and
- congestion control, so that the receiver can adjust its transmission rate according to the network conditions.

In addition to NACK-based repair strategy and feedback suppression, FEC techniques can be used with NORM to further reduce the amount of repair requests and repair transmissions.
Table 5-1: Summary table for Reliable Multicast protocols (Source: [28], [30])

<table>
<thead>
<tr>
<th>Protocol</th>
<th>Reliability Mechanism</th>
</tr>
</thead>
<tbody>
<tr>
<td>MTP/MTP-2</td>
<td>NACK</td>
</tr>
<tr>
<td>MDP/MDPV2</td>
<td>NACK</td>
</tr>
<tr>
<td>AFDP</td>
<td>NACK</td>
</tr>
<tr>
<td>MFTP</td>
<td>NACK, FEC</td>
</tr>
<tr>
<td>R RCMP</td>
<td>NACK, FEC</td>
</tr>
<tr>
<td>NORM</td>
<td>NACK, FEC, feedback suppression, congestion control</td>
</tr>
<tr>
<td>BE</td>
<td>Analysis and application layer</td>
</tr>
</tbody>
</table>

FLUTE

FLUTE [31] acronym stands for File Delivery over Unidirectional Transport. As stated by its name, this multicast protocol is particularly well suited for file transfer. The fact that it is unidirectional makes it particularly interesting for our case; it means that no traffic will go on the Satellite return link. FLUTE is based on two protocols: ALC [32] and LCT [33].

Similarly to NORM, FLUTE protocol implements the following functions:

- Congestion control to avoid flooding the network (CC);
- FEC for the error detection;
- Feedback suppression.

FLUTE provides application level metadata on top of ALC+LCT. It basically provides a means to identify, locate and restore file at the receiver side using FLUTE headers and File Delivery Table (FDT) instances.

5.2.8 Satellite multicast for linear distribution of video programs

As mentioned earlier, satellites use multicast for many applications such as TV distribution and private network services. A satellite multicast approach proposed by SES is described in [34], [35] where the goal is the linear distribution of video programs to end user devices. In this work, video programs received from one or more content sources, are encoded and encapsulated into IP video streams. The streams are transmitted to a satellite operator Network Operation Center (NOC), which encrypts the streams. The encrypted streams are transmitted, as multicast video streams, through a satellite, to one or more cable television headends, and then, over an IP-based network, to cable modems located within subscribers’ premises. An IP-enabled end user device transmits, to a gateway associated with one of the cable modems, a command requesting that the IP-enabled end user device obtains a requested video stream as a unicast stream. Upon receiving the command, the gateway triggers the conversion, at the gateway, of the requested video stream from multicast to unicast for transmission over a home network provided within the subscriber’s premises. The requested video stream is then transmitted, over the home network, to said end user device, which decrypts and decodes the requested video stream into the original video program, so that it can be displayed to the end user. This process is presented in Figure 5-4.
However, this technique is based on the existence of a home network equipment converting multicast back into unicast, an equipment that is not present in mobility. This function is necessary since most mobile devices are not capable of receiving a multicast feed, as it requires a new user device chipset design.

The same issues also face the Multicast Adaptive Bit Rate standards (mABR) [16], [17]. A typical M-ABR system can be thought of as a standard ABR video system, which uses a transparent caching proxy resident in the Gateway. That transparent cache can be filled either via unicast or multicast. This allows the Player to switch seamlessly between less-popular content only available on unicast and popular content available on multicast as it is completely transparent to the Player whether the content is delivered to the Gateway via unicast or multicast. In fact, the system can switch seamlessly between unicast and multicast delivery of the same stream, as any content not delivered by multicast will be retrieved via unicast.

While this technology is referred to as "Multicast Adaptive Bit Rate (M-ABR)", it is important to note that individual multicast streams do not "adapt" their bit rates. Rather, the term is used to refer to the multicast delivery of video segment files to the Gateway which subsequently delivers these segments via HTTP when they are requested by a streaming video Player. Each multicast stream only contains a single bit rate. The pre-filling of the Gateway's cache is expected to result in the reliable receipt of fragments by the Player, such that the Player does not adapt and instead chooses to remain at that bit rate. However, for robustness, the manifest still generally contains reference to other bit rate encodings of the same content stream. These other bit rates can be provided on a separate multicast stream or may be available only via unicast retrieval.

### 5.3 Multicast in 3GPP Networks

In this section some 3GPP Multicast techniques are presented. Firstly a presentation of the 4G eMBMS implementation is made followed by two sections more dedicated to 5G.

#### 5.3.1 4G multicast implementation (eMBMS)

3GPP defines MBMS (Multimedia Multicast/Broadcast Service) as a point-to-multipoint service in which data is transmitted from a single source entity to multiple recipients. Transmitting the same data to multiple recipients allows network resources to be shared [18].

The MBMS bearer service offers two modes:

- Broadcast Mode or;
- Multicast Mode.
For MBMS, 3GPP introduced new definitions [36]. Figure 5-5 graphically shows what the MBMS Service area, MBSFN area and MBSFN area reserved cell represents. A MBMS service area contains at least one MBSFN and the MBSFN may contain or not MBSFN area reserved cells.

![Figure 5-5: MBMS definition [36]](image)

The new MBMS concepts are the following:

- **MBSFN Synchronization Area**: an area of the network where all eNodeBs can be synchronized and perform MBSFN transmissions. MBSFN Synchronization Areas are capable of supporting one or more MBSFN Areas. On a given frequency layer, an eNodeB can only belong to one MBSFN Synchronization Area. MBSFN Synchronization Areas are independent from the definition of MBMS Service Areas.

- **MBSFN Transmission or a transmission in MBSFN mode**: a simulcast transmission technique realized by transmission of identical waveforms at the same time from multiple cells. An MBSFN Transmission from multiple cells within the MBSFN Area is seen as a single transmission by a UE.

- **MBSFN Area**: an MBSFN Area consists of a group of cells within an MBSFN Synchronization Area of a network, which are coordinated to achieve an MBSFN Transmission. Except for the MBSFN Area Reserved Cells, all cells within an MBSFN Area contribute to the MBSFN Transmission and advertise its availability. The UE may only need to consider a subset of the MBSFN areas that are configured, i.e. when it knows which MBSFN area applies for the service(s) it is interested to receive.

- **MBSFN Area Reserved Cell**: A cell within a MBSFN Area which does not contribute to the MBSFN Transmission. The cell may be allowed to transmit for other services but at restricted power on the resource allocated for the MBSFN transmission.

- **Synchronization Sequence**: Each SYNC PDU contains a time stamp which indicates the start time of the synchronization sequence. For an MBMS service, each synchronization sequence has the same duration which is configured in the BM-SC and the MCE.

- **Synchronization Period**: The synchronization period provides the time reference for the indication of the start time of each synchronization sequence. The time stamp which is provided in each SYNC PDU is a relative value which refers to the start time of the synchronization period. The duration of the synchronization period is configurable.

The latest capabilities introduced by 3GPP in relation to MBMS services where done in Rel.14 [37]:

- **Support of larger Inter-Site Distance (ISD) at high spectral efficiency** covering up to 15 km.

- **Dedicated or mixed eMBMS carrier** allowing for mixed unicast and broadcast services over a single carrier using up to 100% broadcast resource allocation.

- **New subframe type** without unicast control region to reduce overhead compared with previous eMBMS transmissions.

- **Shared eMBMS broadcast** where operators can aggregate their eMBMS networks into a shared eMBMS content distribution platform.
In the same 3GPP Rel.14, new eMBMS system architecture and media formats were defined:

- **Receive-only mode (ROM):** allows devices without a SIM be subscribed to a subset of services like TV over a cellular network.
- **Free-to-air services:** are related to receive-only mode and allow to use the mobile network for multiple broadcast purposes since the broadcasted data is accessible to anyone including the ones that are not mobile subscribers.
- **xMB interface:** simplifies the access to eMBMS system functionalities to content providers and broadcasters. With this new interface, it is possible to establish TV services through the standardized xMB.
- **A new eMBS Application Programming Interface (API):** primarily for developers of web and user applications to simplify access to complex eMBMS procedures.

The MBMS is defined in a way that creates an end-to-end service between the core network and the end user. 3GPP specifies a new architecture to support MBMS that affects the core network and all the layers in the RAN protocol stack including RRC, RLC, MAC and PHY. The new components with the new interfaces are represented in Figure 5-6.

![Figure 5-6: eMBMS architecture deployment](image)

The MBMS architecture defines new components and requires multiple upgrades on existing components. The new components and their functionalities are:

- **Broadcast-Multicast Service Centre (BM-SC):** The BM-SC provides functions for MBMS user service provisioning and delivery. It may serve as an entry point for content provider MBMS transmissions, used to authorize and initiate MBMS Bearer Services within the PLMN and can be used to schedule and deliver MBMS transmissions. The BM-SC is a functional entity, which must exist for each MBMS User Service. The BM-SC consists of the following sub-functions: Membership function; Session and Transmission function; Proxy and Transport function; Service Announcement function; Security function; Content synchronization for MBMS in E-UTRAN for broadcast mode; Header compression for Mission Critical services using MBMS in E-UTRAN. Figure 5-7 shows the U-plane architecture of the MBMS content synchronization.
• **MBMS-GW**: The MBMS GW is a logical entity that is present between the BM-SC and eNBs whose principal functions is the sending/broadcasting of MBMS packets to each eNB transmitting the service. The MBMS-GW uses IP Multicast as the means of forwarding MBMS user data to the eNB. The MBMS-GW performs MBMS Session Control Signaling (Session start/update/stop) towards the E-UTRAN via MME [36].

• **Multi-cell/multicast Coordination Entity (MCE)**: The MCE is a logical entity with multiple functions (C-plane protocol architecture represented in Figure 5-8).
  - the admission control and the allocation of the radio resources used by all eNBs in the MBSFN area for multi-cell MBMS transmissions using MBSFN operation. The MCE decides not to establish the radio bearer(s) of the new MBMS service(s) if the radio resources are not sufficient for the corresponding MBMS service(s) or may pre-empt radio resources from other radio bearer(s) of ongoing MBMS service(s) according to ARP. Besides allocation of the time/ frequency radio resources this also includes deciding the further details of the radio configuration e.g. the modulation and coding scheme.
  - counting and acquisition of counting results for MBMS service(s).
  - resumption of MBMS session(s) within MBSFN area(s);
  - suspension of MBMS session(s) within MBSFN area(s).

The two different MBMS modes previously described, multicast and broadcast, have different procedures to provide the service. The multicast mode includes UE specific procedures and core network procedures while broadcast mode defines only the procedures related with the core network. It has to be noted that core procedures are the same for multicast and broadcast services. Multicast procedures are the following: subscription (UE specific), service announcement, joining (UE specific), session start, MBMS notification, data transfer, session stop and leaving (UE specific). The broadcast mode keeps the service announcement, session start, MBMS notification, data transfer and session
stop. An example of the timeline for the multicast process with two different UEs is described in Figure 5-9.

Independently of the MBMS transmission mode (multicast/broadcast), the MBMS transmissions may be done in a single cell or in a multi cell transmission. For single cell transmission, the MBMS services are transmitted in the coverage of a single cell. On the other hand, in multi cell transmissions synchronous transmissions of MBMS are done within its MBSFN area. These transmissions are directly related with the MBMS cell type. 3GPP specifies three different cell types based on its capabilities (more details in [36] section 15.3):

- **MBMS-dedicated cell**: cells performing only MBMS transmissions are referred to as MBMS-dedicated cells. UEs not supporting FeMBMS are not supported on these cells. Paging is not supported on an MBMS-dedicated cell.
- **MBMS/Unicast-mixed cell**: cells performing both MBMS and unicast transmissions are referred to as MBMS/Unicast-mixed cells.
- **FeMBMS/Unicast-mixed cell**: a FeMBMS/Unicast-mixed cell is an MBMS/Unicast-mixed cell. UEs not supporting FeMBMS are not supported on these cells and camping of such UEs is prevented by using cell barring mechanism of SIB1. Paging for incoming calls is not supported on such cells and system information change notification as well as ETWS/CMAS notification is provided with L1 signaling. It operates with at least one of the following:
  - subframes 4 or 9 or both configured as MBSFN subframes;
  - subframes that may not contain unicast control region.

MBMS is designed to operate in a network which supports mobility by definition. Once the MBMS session is established, it is required that this session has continuity across different cells ensuring service continuity. Mobility procedures for MBMS reception allow the UE to start or continue receiving MBMS service(s) via MBSFN or Single Cell Point to Multipoint (SC-PTM) when changing cell(s). For each MBMS service provided using SC-PTM, E-UTRAN indicates in the Single Cell Multicast Control Channel (SC-MCCH) the list of neighbor cells providing this MBMS service so that the UE can request unicast reception of the service before changing to a cell not providing the MBMS service using SC-PTM.

For MBSFN transmission, E-UTRAN procedures provide support for service continuity with respect to mobility within the same MBSFN area. Within the same geographic area, MBMS services can be provided on more than one frequency and the frequencies used to provide MBMS services may
change from one geographic area to another within a PLMN. The MBSFN area is static and may be modified by the OAM.

5.3.2 5GxCast Point to Multipoint architecture

In the following section a brief overview of the architectures envisaged by the 5GxCast project is presented. More information on the project can be found at [1], the following sections are extracted from the deliverable D4.1 [38].

5GxCast project focuses on designing a conceptually novel and forward-looking 5G network architecture for large scale immersive media delivery. To do so, two main research pillars are being investigated:

- Develop broadcast and multicast point to multipoint capabilities for 5G
- Design a dynamically adaptable 5G network architecture with layer-independent network interfaces capable of dynamically and seamlessly switching between unicast, multicast and broadcast modes

In the scope of the 5GxCast project, new 5G Network Functions were designed:

5.3.2.1 Xcast User plane Function (XUF)

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 [39], [40]. 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.

5.3.2.2 Xcast Control Plane Function (XCF)

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 [39], [40]. 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.

5.3.2.3 Point to Multipoint Architecture alternatives
5G Reference architecture was derived to introduce those new elements. Three main architecture alternatives were imagined.

5.3.2.3.1 Alternative 1
In this alternative the BMSC as it exists in LTE is 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 interface. The XUF provides multicast and broadcast specific functionalities for the user plane.

The N6 interface and the functionalities of UPF are enhanced to support and handle multicast and/or broadcast 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 5-10: 5GxCast Architecture alternative 1](image)

5.3.2.3.2 Alternative 2
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.
5.3.2.3.3 Alternative 3

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 all standardized functionalities. 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.
5.3.3 Convergence of wireline IP Multicast in 5G

Convergence of wireless and wireline connectivity is under study in 3GPP and their current work is available in 3GPP TR 23.716 [41].

This TR reuses the concept of Residential Gateway (RG) specified in BBF TR-124 [42]. Two main types of RG are identified:

- 5G-RG: 5G Residential Gateway, acts as a UE and interfaces with 5GC through N1 interface;
- FN-RG: Fixed Network Residential Gateway, acts as a UE but does not support N1 signalling
- HA 5G-RG: Hybrid Access 5G Residential Gateway, supports dual connectivity (5G and wireline);

Those RGs may provide 5G Connectivity to UEs. Those UEs can be:

- 5G Capable UEs;
- Non 5G capable UEs.

Residential Gateways can connect to the 5GS either through the NG RAN or through W-5GAN (Wireline 5G Access Network).

A list of possible architectures is presented in the 3GPP document. Presenting exhaustively the TR is not the purpose of this document so we will focus on one of the architecture: 5G-RG connected to 5GS through NG-RAN for 5G capable UEs.
On this figure, the 3GPP UEs have a connection to the 5GS through N1 interface, 5G-RG is connected to the NG-RAN and relays N2/N3 messages to the UEs. The Interface between 5G-RG and NGRAN is yet to be defined.

Using this architecture, a list of solutions for particular use cases is presented. In particular, a solution dedicated to Multicast for IPTV delivery is approached. This IP Multicast scenario can be extended to the work we do in SaT5G and provide the foundations for a converged IP Multicast solution in 5G. This solution has the advantage of not requiring MBMS-like functionality on the End User equipment and might therefore be easier to implement for “generic” software providers (coming from the fixed world).

In this scenario, 5G-RG gateway relays the IGMP join messages from the STB to the elected UPF. The elected UPF then transfers the multicast to the 5G-RG on the PDU session reserved (not pasted here as it is similar to the “normal” 3GPP registration procedure).

This workflow will serve as a basis for section 6.1.4 where we will extend it to match our caching use case.

### 5.4 Multicast for content caching

We previously introduced the concept of content caching for content and media delivery. The main purpose of content caching is to enhance the QoE and the network resource usage. The latter is
extremely important in a mobile environment, network resources are limited and shared by multiple users.

We will here explore content caching in the context of media delivery, in particular Video on Demand (VoD). The concepts remain the same for other use cases (like webpage caching); a cached VoD can be seen as an object stored and retrieved from a cache server, thus what is described here can be extended to meet other use cases’ requirements.

The content caching procedure is not linked to the UE streaming session requests. When a user asks for an asset then this asset is played from the cache if available, otherwise it is streamed from the origin. Having stated this, we can see that several possible implementations for content caching can be envisaged:

- Caching of popular assets during off-peak hours
- Dedicated link where asset to be cached are regularly pushed
- Prefetching on the fly

As described in previous section, on the fly prefetching is based on a unicast connection between the cache server and the origin. Using multicast in this context is not suitable as the prefetching algorithm is not shared between cache servers, one server may decide to prefetch some assets while other may not require this asset at this moment. However, we can see that there is a real interest for the two other cases. It can be envisaged for example that multiple cache servers would share the same list of cached assets (see [8]).

The following figure gives a representation of an arbitrary and simplified CDN topology with several points of presence across the globe.

![Figure 5-15: CDN Topology](image)

This CDN topology is divided into three main regions: Americas, Europe, and Asia. Those regions have different usage models and desired content, thus, we can imagine that the list of assets cached in the respective points of presence (POPs) is different. This has been symbolized by the yellow, red and blue arrows.

It is evident on this figure that multicast has a role to play in this CDN context. Cache servers belonging to the same region may subscribe to the same multicast, this multicast would contain the assets to be cached. On this figure we represented only a few POPs, there is however tens of them in each “region”, using multicast transport will therefore reduce tremendously the bandwidth required for the caching.

The downlink bandwidth offered by the satellite network is particularly suitable in this environment. As the assets are cached in background, there is no requirement in terms of latency (which is the main drawback while using a satellite link) or bitrate. Therefore, the affected resources for the “caching downlink” can be closely controlled by the satellite operator to avoid any impact on the other services.
5.5 **Multicast for 5G VNF update and configuration**

In today's market, business success heavily depends on agility, in the sense that how fast a business can respond to changes, bring new services to the market, and do it all within an IT environment that is simple and inexpensive to operate. In the traditional models, most services were tied to specialized hardware, demanding a long period (in order of weeks or even months) for implementing new services for customers - deploying the appliances, configuring them, provisioning the service. Scaling services was another complex and manual effort which normally used to reduce the organizations competitiveness capacity and forced them to miss business opportunities. To keep the market alive, overprovisioning was a common practice. However, it leads to higher operational costs, with most resources typically sitting idle. For these reasons, service providers and enterprises decided to virtualize network functions to make their environments more agile, scalable, and easier to manage.

With the help of the virtualization technology, it is possible to guarantee:

- **Service agility**, with the ability to dynamically deploy, monitor, and scale Network Services (NSs), so it is possible to on-board new applications faster;
- **Simplified operations**, with an open and extensible architecture that abstracts away NFV complexity and lets service provisioning with reusable data models;
- **Lower OpEx and optimized resource consumption** by automating NS monitoring, scaling, and recovery;
- **Accelerated innovation**, with the ability to integrate with any standards-based NSs, orchestration or assurance system, or custom applications.

According to ETSI MANO, a NS is an end-to-end set implemented by VNFs instantiated on the NFVI, working together towards a single goal. Some examples of a NS would be vEPC, vIMS or vCPE. Each NS would pass several stages during its lifetime, known as lifecycle actions. Figure 5-16 presents generic NS lifecycle actions.

![Figure 5-16: NS lifecycle actions](image)

- **On boarding**: it means on board new VNF type as long as it meets the prerequisites for supporting it in a Virtual Infrastructure Manager (VIM) environment, e.g. OpenStack and VMware. Just to understand the point better, Openstack environment supports QCOW2 image format and config drive support for the VNF bootstrap mechanism.
- **Deploying**: when a VNF is deployed, "day-zero" configurations are applied for a new service. A typical configuration includes credentials, licensing, connectivity information (IP address, gateway), and other static parameters to make the new virtual resource available to the system. Also licenses for the new VNFs are activated.
- **Monitoring**: host hypervisor, whether KVM/OpenStack or VMware, is normally used to monitor the health of virtual machines. It tracks performance metrics such as CPU use, memory consumption, and other core parameters. The requester can specify all of the characteristics (for example, vCPU, memory, disk, monitoring KPIs, and more) typically associated with spinning up and managing a virtual machine.

- **Healing**: VNFs can be healed when there is a failure. The failure scenarios are configured in the KPI section of the VNF data model. KPI are used to monitor the VM and the events are triggered based on the KPI conditions. The actions to be taken for every event that is triggered is configured in the rules section during the deployment.

- **Updating**: starting day zero, deployment might be updated during their lifetime. You can either perform all the updates (that is, add or delete a VNF, add or delete an ephemeral network, and add or delete an interface) in a single deployment or individually.

- **Undeploy**: it is also allowed to undeploy an already deployed VNF. This operation is either done using the northbound APIs or through the portal.

Considering the NS lifecycle events, it seems that multicasting would be beneficial at “on boarding” and “updating” phases. To understand the benefits of multicasting for 5G VNF, let’s have a look again on the multicast definition. Multicast is communication between a single sender and multiple receivers on a network. In this sense, we can assume that on scenario where the VNF developer is far from the point where it is consumed using multicast would save network resources as well as accelerate the service provisioning time. Of course, that includes both introducing a new service to the market as well as updating already available services with some advanced features.

For example, consider a case where VNFs are developed by a Spanish SME and they will be consumed on an in-flight communication and entertainment system provided by a German company. Upon developing the VNF package locally in Spain, SME needs to transfer it to the consumption point, in this case different locations in Germany. Using multicast would provide many benefits as it could optimize resource consumption and provisioning time, keeping in mind that the same VNF package needs to be delivered into the several destinations where it will be offered.
6 SaT5G Caching Architecture

In this context the term “SaT5G Caching Architecture” simply means the caching architecture developed by the SaT5G project to deliver content to the edge in a combined terrestrial satellite 5G network that will be prototyped, validated and demonstrated within the project.

In the following sections, the presented architectures are based on the generic architecture for the integration of satellite links in the 5G network that is presented in deliverable D3.1 [43]. Three new elements were added to the generic architecture:

- A cache equipment within the Edge Network titled “Local DN MEC”. This cache is linked to a local UPF which may be elected by the SMF based on traffic steering policies.
- A link between CDN Server and Local DN. This link may be unicast or multicast and is used to populate the local DN with popular assets.
- A new Caching Application Function in the 5G Core: decides on traffic steering policies and provides them to the SMF via PCF.

The purpose of these elements and how they interact with each other is further described in following sections.

6.1 Caching Workflows

6.1.1 Predictive and analytics-based caching

Predictive and session counting based caching are detailed in the same section because there is a high interdependency between both features. The list of assets that will be pushed to the local cache is determined using the result of both algorithms.

The workflows differ a bit depending on whether we are in a MNO or an externally orchestrated architecture. Both options are described in the following subsections.

In all cases, the predictive caching is managed by the CDN which has the content knowledge. If the CDN envisioned that an event is going to be popular (new series launch, popular TV shows...), then it may request in advance the caching of that specific asset. How the CDN informs of the popular assets is implementation dependent and out of the scope of this document. This will be addressed in WP4.

In the remainder of Section 6 we are presenting figures where the following new logical functions are introduced:

- **Service Router**: function in charge of determining the redirection policies to be applied in the network;
- **Caching Management**: function in charge of electing the assets to be cached locally to the Edge. This may be computed using analytics information and/or CDN preferences;
- **Multicast Controller**: function in charge of pushing the assets in the network using multicast technology.

6.1.2 Caching orchestrated by MNO

6.1.2.1 High level architecture

In this scenario, the MNO is in charge of orchestrating the caching of assets within the Local DN. The Caching Application Function holds all the intelligence: it decides on which media content should be pushed, and how the traffic is steered in the 5G Network. An API is shared with the CDN platform; this API may be used by the CDN to indicate the assets they’d like to push.
6.1.2.2 Sequence diagram
The following diagram depicts the logical functions and their interactions. We afterwards detail those interactions in a sequence diagram.

As previously stated, the MNO orchestrates the caching function. The “AF” has a leg in both Control Plane and User Plane, it holds the following functions:
- Service Router;
- Caching Management;
- Multicast Controller.

The Service Router uses the information from the Control Plane (Subscriber info, Location, Available Bandwidth, etc.) in order to steer the traffic to the right UPF (through PCF and SMF).

The Caching Management function elects and retrieves from the CDN the assets to be pushed to the local cache. This election is based on CDN inputs (through shared API) along with analytics information.

Multicast controller based on information from the Caching Management sends the asset to the local cache through a Carousel based multicast.

The detailed caching workflow diagram is presented in Figure 6-3. There are two main processes that take place which are explained step by step below:

**Multicast carousel management:**

1. The multicast feed which will be joined by the local caches. This multicast is carousel based and contains the assets;
2. Local cache retrieves the assets from the carousel;
3. Local cache stores the missing assets within its cache server.

**Multicast carousel configuration:**

1. The CDN platform calls the API provided by the MNO and gives a list of asset sources to be cached. This list contains only the source URL for the assets;
2. NEF stores this information within the UDR;
3. Periodically the Caching Management function retrieves the list of asset sources that the CDN would like to be cached from the UDR;
4. Caching Management retrieves the list of currently popular assets from the analytics. How the analytics determines what is popular is out of scope (can be based on session counting, user location…);
5. Caching management merges both information (Step 3 and 4) and regenerates the carousel to be sent in the multicast. If there are missing assets in its local storage then it retrieves them from the CDN source and stores them locally;
6. Carousel is provided to the Multicast controller which in turn modifies the multicast with the new carousel.
Figure 6-3: MNO Operated caching: caching workflow
6.1.3 Caching orchestrated externally (e.g. CDNaaS)

6.1.3.1 High level architecture
In this scenario, the MNO is only in charge of steering the traffic to the right UPF. The Caching function is still the CDN responsibility: this allows the CDN to maintain its own caching algorithm. No API is shared between the MNO and the CDN.

![High level architecture diagram](image)

Figure 6-4: Third party caching orchestration: high level architecture

6.1.3.2 Sequence diagram

![Sequence diagram](image)

Figure 6-5: Third party caching orchestration: logical block diagram
The same logical functions appear in this diagram, they are however handled outside the 3GPP
domain except the Service router which is still managed by the MNO. The service router works
similarly to the previous architecture, the UPF election is based on the CP information.

It should be noted that we describe here one possible implementation. Where the CDN places its
caching function is completely implementation-dependent. It could be placed directly on the CDN
platform or externalized or even managed by a third party.

In the following sequence diagram, we assume that the CDN operator holds the caching management
function while the content source is not its responsibility.

The detailed caching workflow diagram is presented in Figure 6-6. Similarly to the workflow in
presented in Section 6.1.1 two main processes take place:

**Multicast carousel management (same as previous option):**

1. The multicast feed which will be joined by the local caches. This multicast is carousel based
   and contains the assets.
2. Local cache retrieves the assets from the carousel.
3. Local cache stores the missing assets within its cache server.

**Multicast carousel configuration:**

1. Periodically, the Caching Management checks from its analytics server the list of popular
   assets.
2. Then it checks internally if some assets are expected to be popular and should therefore be
   cached.
3. Based on this information, a new carousel is generated. If any assets are missing then they
   are retrieved locally from the content source.
4. Carousel is provided to the Multicast controller which in turn modifies the multicast with the
   new carousel.
Figure 6-6: Third party caching orchestration: caching workflow
6.1.4 IP Multicast for caching

In Section 5.3.3 we presented a possible way to implement IP Multicast in 5G Network. Using this idea, we extend the concept here by modifying the architectures proposed in the TR and applying it to our local caching use case.

As we have seen in the previous section, we use the satellite multicast capabilities to push content from the CDN to a local cache connected to a local UPF. On this basis, we came up with this architecture:

![Figure 6-7: IP Multicast architecture for content caching](image)

This is based on one of the possible architectures presented in the 3GPP TR. There may be other applicable architectures (like a local cache which would be 5G capable, a FN-RG connected to a W-5GAN …), but for simplicity we limit our discussion to the one shown above.

The local cache on the left is connected to a 5G-RG and receives the multicast. The CDN holds the caching management function which sends the multicast on the network.

For this architecture to work, the following interfaces need to support IP Multicast: N3, N9, N6.

The corresponding workflow is the following figure:

![Figure 6-8: IP Multicast for pre-caching workflow](image)
6.2 Prefetching workflows

Here we consider two different scenarios concerning the working mechanism of prefetching. The key difference between them is which network function is responsible for re-directing UE content requests to the MEC server. In the first scenario, AF provides redirection policy to the DNS server on for which video content objects the corresponding UE requests should be redirected to the MEC. In the 2nd scenario, DNS is not aware of such intelligence, but instead the Edge UPF will be responsible for such redirection upon the informed policy from AF.

Figure 6-9 shows the high-level architecture and workflow for the 1st scenario for DASH/HLS video prefetching at the 5G mobile edge. To start with, the content provider, which can be either any third party or the MNO itself (in case the MNO also takes the role of content provider) should define specific prefetching policies through the Application Function (AF) according to the 5G network architecture. For instance, the most important configuration policy is how many DASH video segments should be prefetched for each content session through the satellite backhaul? Such policies will be actually enforced at the MEC server which is responsible for storing the prefetched or cached video content as the local source. In addition, the mechanism should be also developed to redirect the user content requests to the local MEC rather than the remote content source. A common approach is through the DNS system where AF can reconfigure the local DNS server to point to the IP address of the local MEC which is ready to launch the prefetching operation upon any incoming requests on the content URL. If the user equipment (UE) has received the returned DNS response that points to the local MEC server, then it will establish a secured TLS tunnel with the MEC server to start content delivery. In case the content has not yet been locally cached, the MEC server will also establish a secured TLS connection with the remote content source for prefetching purpose. In the context of the DASH applications, the UE will first request to the MEC server the manifest file known as MPD (media presentation description). The MEC server will on-demand fetch the MPD from the source and return it back to the UE. Upon receiving the first request on the DASH video, the MEC server will start to activate the prefetching operation which makes its segment downloading progress constantly ahead of the UE request.

(a)
Figure 6-10 shows the high-level architecture and workflow for the 2nd scenario. In this case AF informs the PCF on the list of content items for which their UE request should be redirected to the local MEC server. PCF will create the corresponding redirection policy at the local Edge UPF which in this case becomes content aware. Since in this scenario DNS always returns back the IP address of the original content source, the UE will attempt to establish connection with the remote source. If MEC is supposed to provide prefetching service for this content, then Edge UPF will perform the local redirection (e.g. ARP based redirection) towards the local MEC server. As such the UE will establish a secured TLS tunnel with the MEC server to start content delivery. In case the content has not yet been locally cached, the MEC server will also establish a secured TLS connection with the remote content source for prefetching purpose. In the context of the DASH applications, the UE will first request to the MEC server the manifest file known as MPD (media presentation description). The MEC server will on-demand fetch the MDP from the source and return it back to the UE. Upon receiving the first request on the DASH video, the MEC server will start to activate the prefetching operation which makes its segment downloading progress constantly ahead of the UE request.
The basic prefetching operations outlined above do not consider user mobility. In case an end user on the move has a handover between base stations that are attached to different MEC servers, then there is a requirement of coordination between the MEC servers in order to provide seamless service continuity. In the literature there have been works on achieving similar objectives in different environments like mobility support for Named Data Network (NDN), and the algorithm there can be directly adapted and used in the SaT5G environment.

### 6.2.1 Enhancing prefetching workflow with MPEG DASH SAND

The MEC server may be implemented as a DANE (see Section 3.3.2), in which case the DASH clients and the DANE should support the MPEG DASH SAND protocol. Via the SAND protocol, the DASH clients will be able to provide the MEC DANE with relevant information to feed the fetching logic of the MEC DANE, such as which segments are expected to be requested next and by which time they should be delivered. This information may be of particular use in the case in which dual (terrestrial and satellite) backhaul link is available, where then the MEC DANE can choose to fetch some segments via the terrestrial link and others via the satellite. Figure 6-11 shows an example of prefetching using this mechanism. Based on the information received by the DASH client, the MEC DANE decides to prefetch some segments (segment 2 in the Figure 6-11) via the terrestrial link, while some other segments (segments 3 to k in the Figure 6-11) from the satellite link.
Figure 6-11: Prefetching architecture and workflow with SAND (Scenario 1) Architecture for live content delivery

6.3 SaT5G mABR Architecture for live delivery

Compared to the generic architecture presented in [43], and based on the presentation of mABR made in section 5.1, the following new elements are added:

- A dedicated AF in charge of redirecting the requests to the UPF connected to Function Y;
- A Function X managed which converts live from unicast to multicast;
- A Local DN acting as function Y (in mABR terminology), this function has joined the multicast to receive the live. The Function Y receives the multicast and sends it on as unicast to the UE.

When a UE requests a session on a popular live, this UE is redirected to a decentralized UPF connected to a Local DN. This Local DN receives the popular lives in Multicast from the 5GC and converts them back in Unicast for the UE. The satellite link is used for its inherent Multicast capabilities and efficiencies, and its ability to reach any point across the globe.

The architecture is very similar to what has been presented for Caching, the workflows are equivalent. The following figure illustrates the case when the multicast is managed by the CDN.
Figure 6-12: SaT5G mABR Architecture

The corresponding workflow is presented in the next figure:

Figure 6-13: mABR SaT5G Workflow

Description:

Live multicast management:

1. The Multicast Controller pulls from the origin the live in unicast;
2. Multicast Controller generates a multicast flow which is joined by the Local DN. Local DN serves the live in unicast.
Live multicast configuration:

1. Periodically, the Live Management checks from its analytics server the list of popular live content;
2. Then it checks internally if some live content is expected to be popular and should therefore be multicast;
3. Based on this information the Live Management configures the Multicast Controller to create a new multicast for those live content.
7 Summary and Conclusions

7.1 Scope

Content caching has been looked at and implemented in terrestrial ISP networks for many years. Today, caching is a broadly accepted technology driven by the seemingly insatiable growth in media consumption. Within mobile networks, contents can be stored at either the EPC or at the RAN (Radio access network) using traditional caching policies such as LRU, LFU, FIFO and random caching. This deliverable has reviewed the state of the art and emerging work to define best practise for incorporating satellite communication links (satcom) into 5G networks.

7.2 Content distribution

Content Delivery Networks (CDNs) are composed of a system of servers allocated on different points of presence (POP) in an operator’s network. The servers are “super-computers” defined by their capacity of storage, processing, and input/output bit rate; they are often implemented to provide IP TV to the operators’ customer base. These servers make the video retrieved from the encoders available in the delivery networks, and replicate the video for every end user.

![Figure 7-1: High level content distribution architectures](image)

Over the top (OTT) is different from IPTV. OTT video content is made available to subscribers directly by the providers (e.g., producers, TV channels, or aggregators) via Web portals and apps on a variety of devices, such as computers, tablets, smartphones, and connected TVs. Hence, OTT is more of a multiscreen approach. With OTT, there is no longer a need to use the operator’s set-top box to view content. However, with OTT content delivery, the operator loses the control on how its content is distributed. While operators can guarantee bandwidth dedicated to their IPTV services, OTT video content is processed on the open Internet, similar to any other data. OTT content is transported to the ISP through private CDNs. From this point, nothing can distinguish the video content from the other data in transit in the operator’s network.

Using an adaptive delivery format, an ISP can decrease the network load while improving the user quality of experience. Thanks to this technology, when the network has been saturated to a point where it cannot transport an integral signal, the receiver requests a lower bit rate from the server, reducing the amount of data so that the signal can be delivered. Of course, the quality of the video content resulting from this technology is slightly degraded, but this is more acceptable than a break in service. When the network conditions become better, the bit rate and the quality increase again accordingly.

Currently, the most used adaptive formats are Apple HLS (HTTP Live Streaming) and Microsoft HSS (HTTP Smooth Streaming). An open standard called MPEG-DASH (Dynamic Adaptive Streaming Over HTTP) exists, and is also being deployed.
7.3 Content pre-positioning through caches

Content Caching is a way to optimize content delivery. This is typically done by storing assets in various network places, closer to the end user thus optimizing both QoS and network resources. Caching can be triggered from various criteria: session counting, localized content, content popularity or any service specific trigger. One can divide caching in three main algorithms:

- Generic caching based on session counting;
- Predictive caching (prefetching of content before they are requested);
- Prefetching of segments based on current sessions.

Predictive caching is where, based on statistical models, the aim is to predict what/how/when/where a user (or a panel of users) will 'play' in order to cache (or a part of it) before the user hits play [7]. As the content is cached beforehand, the quality of the cached content can be higher since it is made in the background. Such an approach is typically useful for content like TV series or episodes. For example, we can assume that if a user has been watching a particular series it is very likely that this user will play the next unwatched episode. In this case the content of the next episode can be proactively cached before the user starts to consume it. In the context of SaT5G, in case a piece of content is expected to be popular at multiple network locations, the leverage on satellite backhaul links to multicast the content to the local caches to be proactively cached at those edge locations is a promising solution.

Satellite links can be directly involved in the real-time delivery of video content in 5G environments. It is apparent that the long propagation delay of satellite links is a significant issue that can affect the end-to-end performance of DASH (Dynamic Adaptive Streaming over HTTP) / HLS (HTTP Live Streaming) video streaming based on TCP.

MPEG DASH SAND (Server and Network-Assisted DASH) is an MPEG/ISO standard (ISO/IEC 23009-5:2017) which has been developed to enable consistent and higher quality service for adaptive video streaming applications using DASH delivery. To achieve this goal, SAND establishes an asynchronous communication channel between the DASH client and a DASH-Aware Network Element (DANE), through which the DASH client obtains assistance from the streaming server and/or the network. DASH clients can inform the MEC DANE about the next segments to be requested by sending the SAND status message “AnticipatedRequests”. This way, the MEC DANE may decide to pre-fetch the content for faster delivery. The client may also provide a deadline for specific segments, such that the MEC DANE can schedule and pace the prefetching of the segments.

Within the context of a dual 5G backhaul network connection (i.e. one via satellite), the MEC DANE may also choose to pre-fetch some segments via the terrestrial link (i.e. if they are needed soon by the requesting DASH client), and the bulk of other segments via the higher bandwidth satellite link (e.g. when their due time is far enough). In this way, the load on the backhaul networks can be balanced while at the same time providing high quality streams to the DASH clients.

7.4 Caching within ETSI

Multi-access Edge Computing (MEC) offers application developers and content providers cloud-computing capabilities and an IT service environment at the edge of the mobile network. This environment is characterized by ultra-low latency and high bandwidth as well as real-time access to radio network information that can be leveraged by applications. MEC provides a new ecosystem and value chain. Operators can create new revenue streams by exposing their Radio Access Network (RAN) edge to authorized third-parties for application hosting and providing RAN-based services, allowing them to flexibly and rapidly deploy innovative applications and services towards mobile subscribers, enterprises and vertical segments.

The MEC initiative [11] is an Industry Specification Group (ISG) within ETSI. The purpose of the ISG is to create a standardized, open environment which will allow the efficient and seamless integration of applications from vendors, service providers, and third-parties across multi-vendor Multi-access Edge Computing platforms.

In the frame of satcom, deploying applications and delivering services at the edge of the satellite network can bring even more benefit to the satellite end user as the impact on the delay, and subsequently the user’s QoE bring greater benefit.
ETSI MEC defines a framework for hosting and delivering services at the network edge. The ETSI MEC framework provides a secure and trusted platform at the network edge, within the trusted domain of the telecom’s operator, to host and provide services to the end user and provide additional services that are unique to the network edge (e.g. real time location, congestion information).

Figure 7-2: Edge computing [11]

Device-to-Device communication was introduced in LTE REL12 and enables devices located within the same cell to directly communicate over allocated spectrum. The standardization of D2D is not complete yet; however, it may offer a very simple and standard option to deploy caching services and it turn out a particularly well-suited technique to be married with satellite content update capabilities. A simplified overview of Device-to-device communications is:

- “Serving Devices” can register their service and publish the service they offer using multicast;
- “Client Devices” that wish to use the service ask the cellular base station to allocate a slice of spectrum (time and frequency slot);
- The client device uses the allocated bandwidth to directly communicate with the serving device without utilizing the cellular infrastructure.

Adapting this approach to satcom appears worthy of further investigation however the immaturity of this concept prevents further analysis in SaT5G.

### 7.5 Multicast transport

Nowadays, Over-The-Top content delivery is dominated by the adaptive bit rate (ABR) protocols. Those L4 protocols take advantage of HTTP’s universality:

- Ubiquitous: implemented everywhere: browsers, smartphones, tablets;
- Ease of player development;
- Easy firewall traversal;
- Relies on TCP congestion algorithms.

In the case of mABR (Multicast Adaptive Bit Rate), live channel is multicast within the backbone and converted back to unicast closer to the end user device.
As shown in the figure above, a single multicast connection is managed between multicast server (represented by Transcaster server in this figure) and the Multicast Gateway function within the Home Network.)

This figure is a possible implementation of mABR. mABR is being standardized by the DVB Group [12]. The implementation proposed here does not reflect all the possibilities offered by the DVB standard. As an example, the conversion from unicast to multicast could be done at other points in the end-to-end chain.

The cost benefits of using satellite to distribute multicast content as the number of receiving nodes increase are illustrated in the figure below. The real-world values illustrated will be different but this simply shows that the bandwidth cost for satellite multicast remains fundamentally the same regardless of the number of nodes whereas the terrestrial multicast bandwidth costs are roughly proportional to the number of nodes.

Satellites have for many years regularly used multicast for many applications such as TV distribution and private network solutions. A satellite multicast approach proposed by SES is described in [34], [35] where the goal is the linear distribution of video programs to end user devices. In this work, video programs received from one or more content source, are encoded and encapsulated into IP video streams. The streams are transmitted to a satellite operator Network Operation Center (NOC), which encrypts the streams.

![Figure 7-3: Possible mABR implementation](image)

![Figure 7-4: Illustrating the cost benefits of satellite multicast distribution](image)
The encrypted streams are transmitted, as multicast video streams, through a satellite, to one or more cable television headend, and then, over an IP-based network, to cable modems located within subscribers’ premises.

![Diagram showing the linear distribution of video programs to end user devices through satellite](Source: [34])

However, this technique is based on the existence of a home network equipment converting multicast back into unicast, an equipment that is not present in mobility but could instantiated as a MEC function at the RAN. This function is necessary since most mobile devices are not capable of receiving a multicast feed, as it requires a new user device chipset design.

### 7.6 Satellite caching architectures

Three new elements have been added to the generic architecture described in SaT5G deliverable D3.1:

- A cache equipment within the Edge Network titled “Local DN MEC”. This cache is linked to a local UPF which may be elected by the SMF based on traffic steering policies
- A link between CDN Server and Local DN. This link may be unicast or multicast and is used to populate the local DN with popular assets
- A new Caching Application Function in the 5G Core: decides on traffic steering policies and provides them to the SMF via PCF.

### 7.6.1 Caching orchestrated by the MNO

In the scenario where caching orchestrated by MNO; the MNO is in charge of orchestrating the caching of assets within the Local DN. The Caching Application Function holds all the intelligence: it decides on which media content should be pushed, and how the traffic is steered in the 5G Network. An API is shared with the CDN platform, this API may be used by the CDN to indicate the assets they’d like to push.
7.6.2 Prefetching mechanisms

Consider two further different scenarios concerning the working mechanism of prefetching. The key difference between them is which network function is responsible for re-directing UE content requests to the MEC server. In the first scenario, AF provides redirection policy to the DNS server on for which video content objects the corresponding UE requests should be redirected to the MEC. In the 2nd scenario, DNS is not aware of such intelligence, but instead the Edge UPF will be responsible for such redirection upon the informed policy from AF.
The following figure shows the high-level architecture for the first scenario for DASH/HLS video prefetching at the 5G mobile edge.

![Figure 7-8: Prefetching architecture (Scenario 1)](image)

The next figure shows the high-level architecture for the second case. In this case AF informs the PCF on the list of content items for which their UE request should be redirected to the local MEC server.

![Figure 7-9: Prefetching architecture and workflow (Scenario 2)](image)

The basic prefetching operations outlined above do not consider user mobility. In case an end user on the move has a handover between base stations that are attached to different MEC servers, then there is a requirement of coordination between the MEC servers in order to provide seamless service continuity. In the literature there have been works on achieving similar objectives in different environments like mobility support for Named Data Network (NDN), and the algorithm there can be directly adapted and used in the SaT5G environment.

### 7.6.3 SaT5G mABR Architecture for live delivery

Compared to the generic architecture presented in [43], the following new elements are added:

- A dedicated AF in charge of redirecting the requests to the UPF connected to Function Y;
- A Function X managed which converts live from unicast to multicast;
A Local DN acting as function Y (in mABR terminology), this function has joined the multicast to receive the live. The Function Y receives the multicast and sends it on as unicast to the UE.

When a UE requests a session on a popular live, this UE is redirected to a decentralized UPF connected to a Local DN. This Local DN receives the popular lives in Multicast from the 5GC and converts them back in Unicast for the UE. The satellite link is used for its inherent Multicast capabilities and efficiencies, and its ability to reach any point across the globe.

The architecture is very similar to what has been presented for Caching, the workflows are equivalent. The following figure illustrates the case when the multicast is managed by the CDN.

Figure 7-10: SaT5G mABR Architecture
8 References


[38] 5GxCast, "Deliverable D4.1 Mobile Core Network," 2018.

[39] 3GPP, “TS 26.346: "Multimedia Broadcast/Multicast Service (MBMS); Protocols and codecs”.

[40] 3GPP, “TS 29.116: "Representational state transfer over xMB reference point between content provider and BM-SC"."


Annex A – Digital Rights Management

Scope

The following analysis was performed in 2013 within the scope of the FP7 BATS project. The content is reproduced below to illustrate the complexities of this area, noting that in five years the subject matter will have evolved significantly.

A1. DRM Solutions

The ability to enable content access anytime, anywhere, from any device is a crucial success factor for any digital content service. In order to meet these consumer demands and achieve success in emerging digital markets, broadcasters, content owners, mobile operators, and service providers must be able to provide a variety of purchasing and consumption models. The foundation behind any flexible pricing and licensing plan is security. The DRM technology has been evolving moving from an initial copy control system to a view management system, trying to override in his evolution the limitations that have been, and also nowadays are, affecting the traditional CAS system:

- Very limited interoperability and coexistence of multiple CAS on one device;
- Very high costs and limitation in the portability of solutions between hardware platforms and OSS.

DRM providers are aware nowadays that is not anymore a problem of protecting “a network” (as the SKY networks for NDS), but more a problem of protecting each single asset, where the DRM solution selection is shifting from the service provider to the content providers, as the big Hollywood Studios and production companies, who still own the content rights. DRM Providers not only nowadays have to provide a rock solid solution difficult to reverse engineer, and easy to update in case compromised, but they have constantly to evolve their product to the way the content is consumed by the users (new media, new platforms, new devices, new social models). Let’s now see what is the current offer from the DRM providers.

A1.1 Microsoft Windows Media DRM

Windows DRM (WDRM) has been initially designed to operate only on the windows media platform, allowing the delivery of the audio and video content over IP network. The key components are:

- Windows Media Rights Manager – for packaging content and issuing licenses;
- Windows Media Format SDK – for building Windows applications;
- Windows Media DRM for portable devices – to support offline playback on portable devices;
- Windows Media DRM for network devices – for streaming protected content to devices attached to the home network.

WDRM is limited to the Microsoft Platforms, but it had a quite large initial deployment. It allows the playback of an asset using only the Microsoft Media Player on a compatible computer, no restriction on the copy of the asset, supporting both streaming and download (depending on the protection policy).

A1.2 Microsoft PlayReady – playready.com

PlayReady has been initially announced in 2007, building on Windows Media, introducing big changes to the previous scheme:

- Concept of domain – group of devices belonging to the same user, which can share the same licenses
- Embedded Licenses – licenses embedded in the container without requiring a separate connection for the license acquisition (not for live)
- Platform Independent – it can be potentially ported on any kind of non-Microsoft portable device.

A1.2.1 PlayReady ND – playready.com
The ND solution has the objective of simplifying the operation in the home environment when multiple devices are used for the content consumption, with particular focus on the live content. ND has got the following Key Aspects:

- Deliver TV to multiple devices lacking of tuners
- Deliver TV to a broad array of client types
- Ensure secure delivery over the home network in a reliable way.

The solution offered is based on the deployment of a local streaming/license server to be deployed in the home network, reducing the bandwidth requirements from/to the head end services and optimizing the media/license traffic in the home network.

**A1.3 OMA DRM – openmobilealliance.org**

The DRM developed by the Open Mobile Alliance (a pool of mobile phone manufacturers, network operators and IT companies); OMA DRM is an open solution with the main objective to set enforced limits on the use and duplication of content by customers (ringtones, music, wallpapers etc.). The initial DRM approach has been lately integrated with two more lines of product: OMA SRM (secure removable media), and OMA SCE (secure content exchange). OMA DRM has been implemented on over 550 models of mobile phones. Most of the ringtones pre-installed on mobile phones have implemented DRM, but not all the Ringtone and content providers for mobile phones are bothered to introduce the technology due to the extreme fragmentation of the market that not always justifies the costs, by the way, OMA DRM is also supported commercial DSTV CA providers as:

- Discreetix
- Irdeto
- NDS
- Viaccess-Orca

**A1.4 Marlin DRM – marlin-community.com**

Marlin is a DRM platform, created by an open standard community. Marlin main point is that interoperability and openness are essential to sustainable commercial success. The Marlin Developer Community mission is to combine the technology, partners and services to enable the creation of interoperable digital content distribution services. Marlin was originally founded in 2006 by Intertrust, Panasonic, Philips, Samsung and Sony. Starting from 2008 Marlin announced the formation of the Marlin Partner Program (MDC) indicating the initial 25 technology partners qualified for the diffusion of the DRM solution. The use of 3rd party technology partners is the main way nowadays DRM solutions are provided.
Figure A-1-1: Marlin DRM operation

The figure shows the interaction required by the client device for the access and acquisition and consumption of protected content. Content acquisition and business logic are completely separated from each other. Marlin started to have a very large growth from 2009 for IPTV in Japan with a portal with more than 1M subscribers, allowing the consumption of DRM protected content from connected TVs, computers and connected devices. The service provided is similar to the USA Hulu. Marlin has been also adopted for the UK platform YouView and by two key bodies in the DRM management:

- UltraViolet (more below)
- IPTV forum

A1.5 Widevine – widewine.com

Widevine is a Google company, providing a multiplatform DRM and video optimization solutions using industry adopted standards (HTML5, HLS, and DASH). Widevine provides an interoperable DRM solution with an already largely diffused solution on over 1 billion devices covering all the ecosystem of the connected devices (phones, tablets, connected TVs, computers), OSS (OSX, Chrome, Linux, Windows). Differently from the other major DRM providers, Widevine provides also full integration allowing to bring the protection to the HW level with agreements with the top STB chipset manufacturers. As Marlin and PlayReady is part of the UltraViolet alliance. Widevine is used by Netflix and Blockbuster, and it has been the first DRM technology embedded in a browser - Chrome. (Widevine media optimizer).

A1.6 Apple Fairplay

Fairplay is an Apple DRM solution for the prevention of the playback of files on unauthorized computers. Fairplay works using encryption of AAC content for the iTunes Store. Fairplay is embedded in Quicktime Multimedia that is used by Apple for decoding and playback of the audio/video files. Every player that supports Quicktime multimedia is able to play Fairplay protected songs. The protection is mainly limited to the Apple products and a very limited number of other manufacturers providing the encoding limited to AAC format, and restricting the possibility of playing the content at the same time on a limited number of authorized devices (5). As iTunes allows to export songs purchased on the iTunes Store in mp3 format and burn CDs, the DRM protection can be easily eliminated providing an easy way for the creation of unauthorized copies.
A2 DRM Convergence and DRM Service Providers

The way DRM has been evolving is quite different from the CAS solution. From the beginning great importance has been given to the interoperability capabilities between the various providers, with the creation of standardization alliances and the birth of DRM aggregators, who are able to provide integrated platforms that support multiple DRM solutions in a single layer. The key reason stays in the fact that DRM has been diffusing initially between the content providers, then as a product for service providers. Software DRM has been initially almost ignored by the major satellite and cable networks due to the financial involvement they have with the major CAS providers. The market evolution and the large expansion DRM had initially on services that were considered not profitable enough to justify the CAS (VOD and catch up TV). When the popularity of Internet TV services as Netflix, Hulu, IPlayer and similar has matched the large networks diffusion, introducing a major risk in the DSTV infrastructure, the CAS providers had to start to support the new technologies on the side of their native technology, the new DRM solutions to allow them an easier integration in their network and the possibility of providing additional content part of their services without the need of radically jeopardizing their infrastructure.

A2.1 DRM Service Providers: Verimatrix, Buy DRM

As DRM has been having a large diffusion between the content provider and media aggregators, for the operators, the need of simple integration of multiple DRM technologies and CAS has become a crucial factor to guarantee a large offer. Operators, on the other side have to accommodate a diversity of receivers over broadcasting and IP networks, regardless of any embedded DRM, supporting fixed and mobile reception over both managed and un-managed networks. This has been pushing the growth for a new category of DRM related business defined as DRM service providers. This new business allow Internet TV providers, DSTV or any other form of content distributors/providers to be able to provide content on their infrastructure that supports different DRM solutions (and also integrated CAS), providing a complete offer to the end users and an unique operator management interface, smoothing all the corners in the content management distribution and business with full support of the connected devices ecosystem and needs. Companies like Verimatrix and BuyDRM provide integrated DRM solutions including PlayReady, Marlin and CENC (Ultraviolet, see below), with additional easy integration with CAS solutions. This approach allows traditional service providers embracing new hybrid solutions to fully support the native content providers DRM without the need of distribution and consumption silos to store petabytes of transcoded media, and have a simple straight forward distribution and media streaming technology that does not frustrate the end users (end users finish to stream content on services like Netflix from the laptop as the same service is not supported by the Cable TV CAS and this moves the end user to slowly abandon the traditional DSTV solution).
A2.2 UltraViolet – (Is CFF the future?)

UltraViolet is an alliance between five of the six major film studios and including the most diffused DRM solutions (PlayReady, Marlin, Widevine, and OMA), retailers, major ISPs and cable TV companies. The objective is to create the first fully interoperable digital rights authentication solution and cloud based licensing system. Ultraviolet adhere to the motto “buy once, play anywhere” allowing the users of digital home entertainment to stream and download purchased content to multiple platforms and devices, allowing users to store digital proof of purchases under one network account to enable the playback of one content that is platform and point of sale agnostic. Up to Six users are allowed per household, with a maximum of 12 shared devices per asset purchased. Centre of the UltraViolet system is the Common File Format (CFF). Once downloaded the CFF asset can be copied or streamed between the registered devices or from the cloud (depending on the asset right scheme). The CFF format will play only on the compatible/registered UV devices. CFF compatible devices will include STB, computers, game consoles, Blu-ray disc players, connected TVs and smartphones. One interesting point is that UV does not regulate the copy of the assets but simply the access rights that are the only thing managed by the service. UV only coordinates and manages the licenses for each account, but not the content itself, the content may be obtained in a different way. The real objective of UV is to create a digital rights locker more than a digital media storage locker. The CFF format, key point in the design, is still under specification (supposed to be released in Aug 2013).

At the current stage, the common file format uses fragmented MPEG-4 container and the basic encryption uses AES keys, which are then protected using each of the DRM required systems. In this way the final player need to support one between the Ultraviolet DRM supported systems. Potentially Ultraviolet got all the right points to become the industry standard, however the market is very volatile.