**3GPP T****SG-RAN WG2 Meeting #113e R2-21xxxx**

**E-Meeting, January 2021**

**Agenda item: 8.x.x**

**Source: Qualcomm (Moderator)**

**Title:** **Email discussion report for [Post112-e][071][MBS] UP Performance**

**Document for: Discussion and Decision**

# Introduction

In RAN2#112e meeting, reliability for Multicast services delivered in RRC\_CONNECTED state was discussed. Following excerpt shows the summary:

**Chairman:** Think that most other functions is not dependent on RLC-AM. Furthermore, the scope of the WI is a bit large for the TU allocation, understand similar to Ericsson that reliability can be achieved with mechanisms other than RLC-AM for PTM (but the cost w.r.t resource usage may be different dep on mechanism). Suggest to assume for now that RLC-AM is not supported for PTM. If it is shown to be needed it can be added, i.e. this can be revisited.

**Working assumption:** RLC-AM for PTM is not supported (can be revisited but it means that proponents of RLC-AM for PTM need to demonstrate the need, to change this).

Following email discussion was agreed to make progress about NR multicast reliability aspects.

* [Post112-e][071][MBS] UP Performance (Qualcomm)

Scope: Based on WID and agreements:

a) The role of PTM vs PTP to achieve wanted UP performance, identify performance requirements. Can also identify other main potential UP performance issues and their related requirements, if any.

b) Identify Tech enablers / insufficiencies / blockers to meet performance requirements and related justifications. Numbers/justifications can be scrutinized by other companies to establish some level of trustworthiness.

In this effort, RAN2 is not expected to evaluate simulation/quant eval numbers in-depth to the level of consensus (numbers already accepted in R1 may be acceptable also in R2 in similar context).

Intended outcome: Report that reflects the discussion, with potentially agreeable proposals / agreeable observations / identified open issues, e.g. whether to confirm the assumption to not support RLC-AM for PTM.

Deadline: Long

In this document, we will collect views from various companies for above email discussion and provide high level summary of proposals for discussion and agreement.

Rapporteur would like to propose following schedule with two phases of discussion:

* Phase 1 (01-06-2021): Companies are invited to provide inputs.
* Phase 2 (01-12-2021): Final report summary and proposals.

# Background

Rel-17 MBS traffic needs to be delivered from a single data source (Application Service Provider or content provider) to possibly many UEs either by using Broadcast or Multicast mode.

Broadcast service is intended to provide service to all UEs in a given service area and for applications which do not require high reliability QoS. Multicast is intended to serve a group of UEs which have joined a multicast service (i.e. not for all UEs). Main goal of Multicast service is to provide same content for a group of UEs in a radio-efficient manner.

Rel-17 MBS system architecture design goal is to support general multicast and broadcast communication services, e.g. transparent IPv4/IPv6 multicast delivery, IPTV, adaptive bit rate (ABR) multicast services, carousel data services, software updates and delivery over wireless, group communications and IoT applications, V2X applications, and public safety etc.

From TR 23.757 [2], section A.3.2, Reference Architecture, in order to support MBS in 5GS user service delivery, two variants modes of operation exist: one for (transparent) Transport Only Mode, and the other for (full) Service Mode.

In Rel-17 MBS system architecture, MBSF entity which provides service layer functionality is optional ([2], see configuration 1 from figure A.3.2-2) and presence of MBSF entity in SA2 system architecture is mainly intended for interworking between R17 NR MBS and LTE eMBMS/SC-PTM system or to provide Service Mode.

For Transport Only Mode (i.e. without MBSF entity, which means no service layer functionality), 5GS typically provide transport-only functionality and all QoS reliability requirement has to be provided by 5G NR RAN.

When application server (AS) requests service from 5GC, the AS provides QoS requirements to 5GC network entities NEF or MBSF and negotiates type of service mode to be provided (i.e. transparent or full service mode). For multicast transparent mode without involving MBSF entity, all QoS reliability requirements have to be handled by NR RAN.

From TR 23.757 [2],clause 8.4, Key Issue #4: QoS level support for Multicast and Broadcast communication services (highlighting added):

*- The network shall support QoS control per MBS session instead of per user.*

*- The network shall support one or multiple QoS flow for a MBS session.*

*- The network may use dedicated QoS flows for multicast sessions in a PDU session if 5GC individually delivery is use to deliver the 5MBS data packet.*

*- The 5G QoS model and parameters as defined in TS 23.501 [2] clause 5.7 also apply to MBS service with the following differences:*

*- Reflective QoS is not applicable;*

*- Wireline access network specific 5G QoS parameters do not apply to MBS services;*

*- Alternative QoS Profile is not applicable;*

*- QoS Notification Control is not applicable;*

*- UE AMBR is not applicable.*

*-    Session-AMBR if provided is enforced at MB-UPF but not communicated to NG-RAN.*

*Editor's note:       Whether Session-AMBR is required in addition to the MBS service data flow bit rate can be determined by operator policy and/agreement with the service provider.*

*- There is support for both GBR and non-GBR MBS flows.*

*- AF provides the MBS session information description including QoS requirements to the 5GC.*

*- The MB-SMF obtains QoS information and configures the MB-UPF accordingly.*

The QoS parameters that are used for multicast QoS Flows do not only include 5QI but also ARP, GFBR, MFBR.

Some applications that can be served by Rel-17 MBS may have QoS requirements for packet delay budget up to 300ms and packet loss reliability up to 10^-6 as shown in the Appendix of this report.

For the full-service layer mode (details are yet to be defined by SA4, but initial decisions are made to re-use basic MBMS user service functionalities as defined in TS 26.346), the following reliability functionalities are available in the MBSF:

* **Application Layer FEC:** Application Layer FEC to be efficient typically needs to span a range of at least 1 second, typically more. This means that the interleaving will add delays of at least 1 second in a very optimistic case, but for efficient usage of application layer FEC, typically several seconds or even minutes of interleaving are preferred. A detailed discussion on the usage and split of AL-FEC and phy FEC for eMBMS is provided in [TR 26.881](https://www.3gpp.org/ftp/Specs/archive/26_series/26.881/26881-f00.zip) [7], clause 9.4 and clause 11.2. As a conclusion it is stated:
  + *There are no AL-FEC Block beginning benefits unless latency is many times the coherence times of the channel in that particular case.*
  + *Application Layer is useful only when latencies of multiple seconds are acceptable.*

While these results are generated for MBMS, it is expected that similar results would be observed for 5G MBS.

* **File Repair:** If files are distributed, typically using some AL-FEC, but some UEs have still not received the file, a repair based on requesting missing byte ranges can be initiated through unicast, using Associated delivery procedures. However, such means are not suitable for real-time services as file repair would only happen after several second, minutes or even hours, i.e. after session completion.
* For unicast, controlling the reliability on application layer can be done by application layer retransmissions. TCP/IP and QUIC (Quick UDP Internet Connection) are providing such means for full reliability, and for latency critical applications (example: VoIP), RTP/RTCP based feedback can be applied. However, MBMS User services do not support application layer retransmission as such technologies have not proven to provide any benefits and are in particular critical to scale as the number of application level retransmission requests may overwhelm network resources and adds significant overhead. Neither TCP/IP nor QUIC nor RTP/RTCP can be supported on multicast.

Overall, MBMS User services do not provide efficient means to reduce the error rates and at the same time maintain latencies below 1 second. Typically, required latencies for efficient distribution are significantly higher than 1 second.

For example, IPTV applications typically require “mean time between failures (MTBF)” of several minutes to hours, and latencies should be below 1-2 seconds in order to meet channel change times and to be on par with other TV distribution systems (satellite, fiber, cable etc).

According to TR 26.925, typical bitrates for such 4K TV services are in the range of 10 Mbit/s. Assuming packet sizes of 1500 byte, we would have 50,000 packets per minute, requiring a 1e-5 loss rates to have an MTBF of 2 minutes. Based on this, 5G RAN has to meet these objectives.

Hence, for such services, the reliability is better taken care of on the radio level , as shown and mentioned above. In eMBMS, only MCS dimensioning is possible (see [TR 26.881](https://www.3gpp.org/ftp/Specs/archive/26_series/26.881/26881-f00.zip) [7], clause 9.4 and clause 11.2), but if the 5G RAN can provide additional means for achieving high reliability, this is very important to meet QoS reliability requirements.

From TR 23.757 [2], clause 4.4

*From the viewpoint of 5G CN, two* ***delivery methods*** *are possible for MBS multicast service:*

*-* ***5GC Individual MBS traffic delivery method:*** *5G CN receives a single copy of MBS data packets and* ***delivers separate copies of those MBS data packets to individual UEs via per-UE PDU sessions****, hence for each such UE one PDU session is required to be associated with a multicast session.*

*-* ***5GC Shared MBS traffic delivery method****: 5G CN receives a single copy of MBS data packets and* ***delivers a single copy of those MBS packets packet to a RAN node****, which then delivers them to one or multiple UEs*

*If 5GC Individual MBS traffic delivery method is supported, a same received single copy of MBS data packets by the CN may be delivered via both 5GC Individual MBS traffic delivery method for some UE(s) and 5GC Shared MBS traffic delivery method for other UEs.*

*From the viewpoint of RAN, (in the case of the shared delivery) two* ***delivery methods*** *are available for the transmission of MBS packet flows over radio:*

*-* ***Point-to-Point (PTP) delivery method****: a RAN node delivers separate copies of MBS data packet over radio to individual UE.*

*-* ***Point-to-Multipoint (PTM) delivery method****: a RAN node delivers a single copy of MBS data packets over radio to a set of UEs.*

*A RAN node may use a combination of PTP/PTM to deliver an MBS packet to UEs.*

*NOTE 2: The PTP and PTM delivery methods are defined in RAN WGs and they are listed here for reference only.*

*As depicted in the following figure, PTP or PTM delivery (with 5GC shared delivery method) and 5GC Individual delivery method may be used at the same time for a multicast MBS session.*



*Figure 4.4‑1: Schematic showing delivery methods*

# Discussion

Companies are requested/encouraged to provide their views for the following questions and explain their answer so that other companies also get opportunity to respond to the detailed views.

1. **Do companies agree that RAN reliability requirements are derived as function of QoS requirements configured by 5GC MB-SMF, and are transparent to type of Multicast application layer transport protocol used?**

|  |  |  |
| --- | --- | --- |
| **Company** | **Agree/Disagree** | **Comments** |
| MediaTek | Agree | We assume that the QoS requirement of a particular multicast service should be ensured below Multicast application layer (i.e. at transportation layer in SA2 language). |
| Samsung | Agree | RAN reliability requirement will be derived by QoS requirements from 5GC. |
| Qualcomm | Agree | RAN has to meet all QoS reliability requirements as requested by MB-SMF based on specific MBS service requirements, and is independent of type of application layer transport protocol used. |
| CATT | Agree | Agree with Samsung. |
| Futurewei | Agree | RAN should be able to provide reliability requested by 5GC MB-SMF. |
| Nokia | Agree | Transparent to the extent that the QoS requirements remain realistic and can actually be met. |
|  |  |  |
|  |  |  |

1. **Do companies agree that, for a certain flow with a given set of values for the QoS parameters (i.e 5QI, ARP, GFBR, MFBR) defined for MBS service as defined in TS 23.757 clause 8.4, multicast QoS requirements are same as unicast QoS requirements with the same values of QoS parameters as specified by 5G QoS model? If not, please provide comments.**

|  |  |  |
| --- | --- | --- |
| **Company** | **Agree/Disagree** | **Comments** |
| MediaTek | Agree | Meanwhile, we think RAN2 may need to present our understanding to SA WGs (e.g. SA2) in order to allow them to extend the current 5G unicast QoS model in the specifications to multicast service also. |
| Samsung | Agree but | “multicast” here is from CN perspective, i.e. MBS traffic to multicast group. It is not about PTM transmission in RAN. SA2 discussion explicitly says that MBS traffic can be served by legacy unicast bearer and RAN2 agreed that gNB decides PTP/PTM. |
| Qualcomm | Agree | SA2 already agreed that QoS characteristics are same for both Multicast and Unicast. |
| CATT | Agree, but | This is only applicable to services supported by multicast.  For services supported by multicast, the multicast QoS requirements should be the same as the QoS requirement of unicast bearer.  But as services supported by multicast is only a subset of services supported by unicast(some services are only supported by unicast,e.g.URLLC), it is natural that the value range of the QoS parameters (i.e 5QI, ARP, GFBR, MFBR) defined for MBS service should also be subset of the value range of the corresponding QoS parameters of unicast. |
| Futurewei | Agree for RAN | RAN should be able to provide similar level of QoS support for QoS flows with the same set of values of QoS parameters (i.e 5QI, ARP, GFBR, MFBR), whether the QoS flow comes to RAN in individual or shared MBS traffic delivery method.  It is 5GC (e.g., MB-SMF) to determine if the same set of values of QoS parameters (i.e 5QI, ARP, GFBR, MFBR) is used between QoS flows in individual or shared MBS traffic delivery method. |
| Nokia | Disagree | In our view, there is no such thing as *multicast* QoS requirements **and** *unicast* QoS requirements: we only have QoS requirements.  Our understanding is that the Rel-15/16 QoS framework applies with QoS requirements being provided per QoS flow (MBS QoS flow in this case). If the RAN chooses to serve the MBS QoS flow with PTM transmissions, it shall ensure that the QoS requirements are met via PTM. However, the RAN may also decide that the QoS requirements are too strict and choose to rely on PTP instead.  Thus which requirements can be fulfilled via PTP or PTM is a RAN decision. |
|  |  |  |
|  |  |  |

From RAN2#111e, following are agreements:

* *Confirm that we will, for multicast services introduce support for PTP and PTM transmission of shared traffic delivered by 5GC, at least for connected mode (this is not intended to exclude other cases)*
* *For a UE, gNB dynamically decides whether to deliver multicast data by PTM or PTP (Shared delivery)*
* *FFS which layer(s) handles reliability (in general), inorder delivery / duplicate handling, and it is FFS how it works at PTM PTP switch.*
* *R2 expect that there may be HARQ with feedback (for PTM) and this is specified by R1.*

From RAN2#112e, following are agreements:

* *For Rel-17, R2 specifies two modes:*

*1: One delivery mode for high QoS (reliability, latency) requirement, to be available in CONNECTED (possibly the UE can switch to other states when there is no data reception TBD)*

*2: One delivery mode for “low” QoS requirement, where the UE can also receive data in INACTIVE/IDLE (details TBD).*

*R2 assumes (for R17) that delivery mode 1 is used only for multicast sessions.*

*R2 assumes that delivery mode 2 is used for broadcast sessions.*

*The applicability of delivery mode 2 to multicast sessions is FFS.*

Rel-17 NR Multicast system is intended to support wide range of services or applications with a wide range of QoS requirements. Some applications like wireless software delivery are required to have high reliability but not delay sensitive and these services need to be delivered in radio efficient manner.

Following are various possible QoS types:

* High reliability, delay sensitive
* High reliability, delay insensitive
* Low reliability, delay sensitive
* Low reliability, delay insensitive

In the rapporteur’s view, the reliability requirement is the same when these applications are delivered either by unicast bearer or multicast bearer.

1. **Do companies agree that any multicast data delivered to UEs either by using UE specific radio bearer or by using multicast radio bearer have to meet the same QoS requirement? If not, please provide comments.**

|  |  |  |
| --- | --- | --- |
| **Company** | **Agree/Disagree** | **Comments** |
| MediaTek | Agree | We did not see the need to differ the QoS support between unicast RB and MRB for a particular multicast service |
| Samsung | Agree, but | We agree Q3. But, it does not mean that multicast radio bearer shall support all the functions of the unicast radio bearer. Which radio bearer is used for the multicast data should be always up to NW. If one type of radio bearer does not meet a specific QoS requirements, the other bearer which can meet the requirements can be configured. |
| Qualcomm | Agree | Same view as MediaTek. For a given multicast radio bearer (MRB), RAN must meet all QoS requirements. |
| CATT | Agree, but | Agree with Samsung.  We think for service with high QoS requirement, PTM only mode could only be used under certain radio conditions (i.e. when the radio condition is above a certain level). |
| Futurewei | Agree | For a given multicast service (of QoS flows determined at 5GC), the same QoS requirement should be met whether it is delivered by UE specific or multicast radio bearer. |
| Nokia | Disagree | We believe we need to be pragmatic and acknowledge that PTM might not (at a reasonable cost) support the same QoS requirements as PTP. Thus, perhaps a more relevant question would be : do companies agree that any QoS requirement that can be met by current DRB shall also be met with PTM leg/transmission of MRB?  In our view, if the RAN receives a BLER QoS requirement of 10-6 from CN then it may not have a choice but to use PTP only. RLC-AM for PTM does not come for free as the overhead of RLC-AM grows linearly. Intuitively, L1 HARQ NACK-only + FEC at higher layers scales better than RLC-AM for PTM.  The overall cost associated with new mechanisms introduced to bring as much reliability to PTM as for PTP need to be first assessed. |
|  |  |  |
|  |  |  |

Depending on Multicast radio bearer architecture, it is possible to provide the reliability at various levels in radio protocol stack. Possible options include

* PHY/MAC based HARQ reliability
* RLC level re-transmission
* PDCP level re-transmission

From layer 2 perspective , it is possible to have the original (first) transmission based on PTM (i.e. to all multicast UEs) and re-transmission either by using PTP (i.e. for a specific UE) or PTM as network implementation decision at each radio protocol level (i.e. either RLC or PDCP layer depending on specific reliability solution).

During RAN1#103, following are agreements for PTP vs PTM from HARQ perspective (highlighting added):

* ***PTP transmission****: For RRC\_CONNECTED UEs, use UE-specific PDCCH with CRC scrambled by UE-specific RNTI (e.g., C-RNTI) to schedule UE-specific PDSCH which is scrambled with the same UE-specific RNTI.*
* ***PTM transmission scheme 1****: For RRC\_CONNECTED UEs in the same MBS group, use group-common PDCCH with CRC scrambled by group-common RNTI to schedule group-common PDSCH which is scrambled with the same group-common RNTI. This scheme can also be called group-common PDCCH based group scheduling scheme.*
* ***PTM transmission scheme 2****: For RRC\_CONNECTED UEs in the same MBS group, use UE-specific PDCCH with CRC scrambled by UE-specific RNTI (e.g., C-RNTI) to schedule group-common PDSCH which is scrambled with group-common RNTI. This scheme can also be called UE-specific PDCCH based group scheduling scheme.*

*For RRC\_CONNECTED UEs, if initial transmission for multicast is based on PTM transmission scheme 1, at least support retransmission(s) can use PTM transmission scheme 1.*

* *FFS: whether to support PTP transmission for retransmission(s).*
* *FFS: whether to support PTM transmission scheme 2 for retransmission(s).*
* *FFS: How to indicate the association between PTM scheme 1 and PTP transmitting the same TB.*
* *FFS: If multiple retransmission schemes are supported, then can different retransmission schemes be supported simultaneously for different UEs in the same group?*

1. **Do companies agree that it is possible to have retransmissions (for both PTP and PTM modes) in MAC (HARQ), transmissions for HARQ, RLC, and/or PDCP depending on multicast radio bearer architecture and retransmission solution support?**

|  |  |  |
| --- | --- | --- |
| **Company** | **Agree/Disagree** | **Comments** |
| MediaTek | Agree | This discussion may be related to the model of the MRB. However, in general, it should be possible to have both PTP based retransmission and PTM based retransmission after PTM based initial transmission and it may occur at both L1 and L2. Such decision should be made by the network at each radio protocol level. The UE reception behaviour may need to adapt to such decision via specified method. |
| Samsung | Disagree | MBS is about “downlink” transmission. Whether to retransmit any downlink packet at any layer is up to NW, without feedback. It’s not prohibited by RAN specifications. But we agree that supporting feedback mechanism depends on multicast radio bearer architecture. |
| Qualcomm | Agree | Based on MRB architecture, it should be possible to support re-transmissions at L1 and L2 level. The question is how to give the NW flexibility to schedule efficient retransmission with feedback, rather than to discuss whether to allow retransmission without feedback. L2 re-transmissions should be based on L2 feedback, since L1 HARQ feedback is not reliable in all cases. |
| CATT | Disagree | HARQ retransmission for both PTM and PTP is necessary, but RLC retransmission for PTM is not necessary, considering the design complexity.  We think the goal of MBS design is to meet the QoS requirement by providing high radio efficiency with design complexity under control. So for service with high QoS requirement , PTM only mode should be used in good radio condition, and switch to PTP to secure the QoS requirement by PTP when radio conditions is bad.  We do not think there is dependency between whether RLC retransmission for PTM is needed and multicast radio bearer architecture. |
| Futurewei | Agree | It is possible to have retransmission in L2 for PTP and PTM modes, at least for the cases where retransmission is done in PTP mode.  Proper design of MBS radio bearer may achieve similar complexity and higher efficiency than using UE specific bearer for multicast service. |
| Nokia |  | Is the question asking whether HARQ retransmissions are supported for both PTP and PTM modes? If so this is a RAN1 issue RAN1 should discuss whether HARQ retransmissions scheduled to a single UE provide benefits or not, not RAN2.  How does the latter part “transmissions for HARQ, RLC, etc.” relate to former part of the question?  If the question is that something more than HARQ is needed on PTM to provide a reliability comparable to PTP, then the obvious answer is yes. But this cannot be decoupled from the associated burdens. |
|  |  |  |
|  |  |  |

As discussed above, one of the key design goals of multicast service is to provide the same level of reliability as that of unicast radio bearer while using common radio resources for all multicast UEs to achieve high radio resource efficiency. Multicast radio bearer reliability can be provided by using PHY/MAC HARQ and L2 reliability.

From TS 38.104 [5], current requirement for DTX -> ACK mis detection is as follows for various PUCCH formats:

8.3.1.2  Minimum requirement

The DTX to ACK probability shall not exceed 1% for all PUCCH formats carrying ACK/NACK bits:



i.e. maximum DTX to ACK false alarm rate = 1%

If we assume the probability of DTX event is 1% = 10^{-2} (i.e. probability of UE not decoding PDCCH), then the probability of DTX to ACK false alarm event is 10^{-4}.

Current requirement for NAK->ACK mis-detection requirements are:

For PUCCH format 0:

8.3.2.2 Minimum requirements

The ACK missed detection probability shall not exceed 1% at the SNR given in table 8.3.2.2-1 and in table 8.3.2.2-2.

For PUCCH format 1:

8.3.3.1.2 Minimum requirements

The NACK to ACK probability shall not exceed 0.1% at the SNR given in table 8.3.3.1.2-1 and table 8.3.3.1.2-2.

For multi-slot PUCCH:

8.3.7.2.1.2 Minimum requirements

The multi-slot NACK to ACK probability shall not exceed 0.1% at the SNR given in table 8.3.7.2.1.2-1.

From above requirements it is clear that depending on PUCCH format, maximum of 0.1% or 1% of the NAK’s will be wrongly interpreted as ACKs.

PHY/MAC based HARQ is not always reliable because NACK feedback can be falsely interpreted as ACK due to radio channel impairments and is more of an issue specifically for FR2. Thus, we can only achieve limited level of QoS reliability using PHY/MAC based HARQ, which is much lower than reliability requirement of 10^-6.

L2 based re-transmission may be needed to support high reliability requirement for many QoS flow applications, which are NOT delay sensitive. Note that HARQ and L2 based reliability are complementary to each other. HARQ alone or L2 alone cannot meet all reliability requirements and both are required to provide high reliable for multicast service delivery to mimic unicast radio bearer level of service delivery.

Without L2 feedback and re-transmission mechanism, higher layer re-transmissions are needed and is very inefficient and adds more delay than L2 based re-tx.

1. **Do companies agree that L1 HARQ alone cannot meet high quality QoS reliability requirements? If not please provide justification.**

|  |  |  |
| --- | --- | --- |
| **Company** | **Agree/Disagree** | **Comments** |
| MediaTek | Agree | Pure L1 HARQ solution cannot meet the QoS requirement for unicast service. That should be the reason for other layers (other than L1) to support their layer-specific feedback and re-transmission mechanism (e.g. at L2) |
| Samsung | Disagree | 1. HARQ retransmission without HARQ feedback (e.g. bundling) can be used for the high reliability. Then “to ACK” error is not a consideration at all.  2. Even without the bundling, the total loss probability of HARQ depends on not only toNACK error but also BLER. The total loss probability can be approximately BLER \* Pr(to ACK error). Assuming Pr(to ACK error)=0.01, 10^-6 can be met by BLER=0.0001. How to set the BLER is fully up to NW implementation and such small BLER can be achieved by robust MCS. |
| Qualcomm | Agree | L1 HARQ has limited reliability and can not meet all QoS reliability requirements. Operating at low MCS means poor spectral efficiency, which may have severe negative impact on other legacy transmission sharing same spectrum with multicast. Multicast key goal is to have same reliability as unicast and operate in radio efficient manner. For NR unicast, L2 retransmission/feedback is supported even when both HARQ and PDSCH slot aggregation features are available. The same criterion is applicable for Multicast as well.  There is identified limitation on L1 reliability, e.g., NACK to ACK error for a given PDSCH BLER target and DTX to ACK for a given PDCCH BLER target, clearly defined in 38.104.  The proper way of loss probability can be calculated as P(DTX)\*P(DTX->ACK)+P(NACK)\*P(NACK->ACK)~=2\*10^-4,  where P(DTX)=0.01 (1% PDCCH target BLER), P(DTX->ACK)=0.01, P(NACK)=0.1 (10% PDSCH target BLER) and P(NACK->ACK)=0.001. It would be impossible to meet high QoS reliability, such as 99.99999%.  It’s too costly in terms of physical radio resources to meet extremely low BLER targets and latency by purely relying on L1 HARQ retransmission and/or L1 repetition. Note that slot aggregation is mainly intended for coverage enhancements and reduce PDCCH scheduling overhead, where the gain of repetition across consecutive slots is limited as well. Any residual BLER after HARQ has to be taken care by L2 reliability. |
| CATT | Disagree | As we commented in Q3, firstly we should clarify on the precondition that for service with high QoS requirement, PTM only mode is only used in certain radio conditions (i.e. when the radio condition is above a certain level).  We think L1 HARQ alone can meet high quality QoS reliability requirements in good radio condition.  Switching to PTP could be a basic solution to secure the QoS reliability when radio conditions are bad. |
| Futurewei | Agree | Both L1 and L2 mechanisms have been specified and applied in LTE and NR for over-the-air transmission, so that reliability can be provided together with satisfactory spectral efficiency.  Only relying on L1 or HARQ retransmission to meet high reliability requirement would put significant strain on radio resources. |
| Nokia | Agree | This does not necessarily imply that something else than HARQ is needed for PTM though (as the RAN is free to choose between PTP and PTM depending on the QoS requirements). |
|  |  |  |
|  |  |  |

TR 38.824[6], section 5.2 provides baseline performance achievable with Release 15 URLLC considering the prioritized URLLC use cases like electrical power distribution, factory automation, transport industry etc.

As part of R15 NR URLLC, key enhancements include DL/UL slot aggregation, DL data channel pre-emption, UL grant free transmission, DL pre-emption, flexible time domain allocation and L2 PDCP duplication with 2 legs in association with DC/CA.

In R16 eURLLC/IIoT, key layer 1 enhancements include inter-UE enhancements, UCI enhancements, PDCCH and PUSCH enhancements, SPS enhancements and CG-PUSCH enhancements. Key layer 2 enhancements include PDCP duplication upto 4 legs using DC/CA configuration, enhanced logical channel prioritization, EHC etc.

Key design goal of URLLC was to meet low latency requirements (up to 15ms e2e delay) and reliability requirements for small payload size but at the expense of **high radio resource overhead.**

**One key design goal of Multicast design is to provide high radio efficiency.** The MBS transmission may havelarge payload size and low delay sensitivity. So, the L1 techniques and L2 PDCP duplication customized for IIoT/URLLC are not appropriate for MBS services. Solely relying on IIoT/URLLC solutions cannot meet high reliability requirements (for delay tolerant applications) and achieve the goal of high radio resource efficiency.

1. **Do companies agree that Multicast key design goal is to provide radio efficiency for diverse applications (variable payload size) with various QoS requirements, which is different than IIoT/URLLC?**

|  |  |  |
| --- | --- | --- |
| **Company** | **Agree/Disagree** | **Comments** |
| MediaTek | Agree | We assume that the IIoT/URLLC type QoS requirement for multicast service is out of the scope of the Rel-17 MBS WI. We also assume that the focus of Rel-17 reliable multicast service should be mainly an enhancement based on the eMBB solution. |
| Samsung | Agree, but | We think providing radio efficiency is important. But it does not mean that any specific mechanism, e.g. RLC AM, is necessary. PTM itself already saves radio resources very much, by reducing duplicate transmissions.  Also, we do not think that MBS requires further latency/reliability enhancements beyond IIoT/URLLC which is considered as an extreme case of latency/reliability. |
| Qualcomm | Agree | URLLC/IIoT requirements are completely different (i.e both reliability and very low OTA latency has to be achieved even at the expense of high radio resource utilization). Most of URLLC/IIoT design enhancements are focusing on fast L1 response at the price of radio efficiency. We agree with MTK that Multicast design goal is to provide high radio efficiency with QoS requirements more similar with unicast eMBB solution. |
| CATT | Partial agree | High radio efficiency is not the only thing to be considered for MBS design.  The design of MBS should consider meeting the QoS requirement by providing high radio efficiency with design complexity under control. |
| Futurewei | Agree | The target use cases in this WID for multicast is different from those in IIoT/URLLC. The required reliability should be achieved together with high radio efficiency by taking advantage of possible PTM transmission opportunities. |
| Nokia | Agree |  |
|  |  |  |
|  |  |  |

1. **Do companies agree that to meet high reliability for delay tolerant Multicast QoS requirements, both L1 HARQ and L2 (RLC and/or PDCP) re-transmission support are needed to meet the goal of radio efficiency for various payload sizes? If companies disagree, please provide explanation.**

|  |  |  |
| --- | --- | --- |
| **Company** | **Agree/Disagree** | **Comments** |
| MediaTek | Agree | Different layers have the specific handling to ensure the reliability of the unicast service. The same paradigm should be applicable to multicast service. Then both L1 and L2 re-transmission should be supported depending the requirement of the QoS for a service. |
| Samsung | Disagree | - L1 HARQ can support sufficiently high reliability by robust MCS and retransmissions.  - Bundling can increase reliability dramatically without HARQ feedback error.  - If NW thinks L2 ARQ is needed, PTP RLC can be used. There is no critical reason why RLC AM is needed for PTM RLC.  - RAN2 already agreed the split-like bearer structure. NW can performs duplicate transmission via both PTP RLC and PTM RLC.  - Even without PTP RLC, gNB can transmit duplicate packets for PTM RLC for additional reliability without feedback.  - RLC AM for PTM is totally new. Considering the scope of this WI, it would be better not to repeat the discussion. |
| Qualcomm | Agree | As we commented in Q5 response, L1 HARQ has limited reliability and slot aggregation is more of coverage enhancement technique. In order to provide spectral efficient Multicast transmission, (repetition is blind and is spectrally inefficient), both L1 and L2 based re-transmission to be supported. Whether L2 re-transmission happens at RLC and/or PDCP is based on specific solution which can be discussed further later. |
| CATT | Disagree | RLC retransmission is not essential for PTM only mode. To meet the QoS requirement of MBS, The design should consider providing high radio efficiency with design complexity under control.  So for service with high QoS requirement, PTM only mode could only be used in certain radio conditions (i.e. when the radio condition is above a certain level), and switch to PTP to secure the QoS requirement when radio conditions is bad. |
| Futurewei | Agree | The required reliability for multicast service should be achieved together with high radio efficiency by taking advantage of possible PTM transmission opportunities. Only relying on retransmission without L2 feedback to meet high reliability requirement would put significant strain on radio resources.  Given that there is concern on the complexity, analysis should be done on the complexity and benefit of supporting L2 retransmission for PTM. |
| Nokia | Agree | This does not necessarily imply that something else than HARQ is needed for PTM though (as the RAN is free to choose between PTP and PTM depending on the QoS requirements). |
|  |  |  |
|  |  |  |
|  |  |  |

# 4. Summary

Based on above discussion, below is summary proposals for discussion and agreement.

[To be update later]

# 5. References

[1] RP-201038: NR Multicast and Broadcast Services

[2] TR 23.757 Study on architectural enhancements for 5G multicast-broadcast services

[3] TR 26.802 Multicast Architecture Enhancement for 5G Media Streaming

[4] 3GPP TS 23.501: "System architecture for the 5G System (5GS)".

[5] TS 38.104 Base Station (BS) radio transmission and reception

[6] TR 38.824 Study on physical layer enhancements for NR ultra-reliable and low latency case (URLLC)

[7] TR 26.881 Forward Error Correction (FEC) for Mission Critical Services.

# Appendix

# From TS 23.501 [4]

### 5.7.4 Standardized 5QI to QoS characteristics mapping

Standardized 5QI values are specified for services that are assumed to be frequently used and thus benefit from optimized signalling by using standardized QoS characteristics. Dynamically assigned 5QI values (which require a signalling of QoS characteristics as part of the QoS profile) can be used for services for which standardized 5QI values are not defined. The one-to-one mapping of standardized 5QI values to 5G QoS characteristics is specified in table 5.7.4-1.

Table 5.7.4-1: Standardized 5QI to QoS characteristics mapping

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 5QI  Value | Resource Type | Default Priority Level | Packet Delay Budget  (NOTE 3) | Packet Error  Rate | Default Maximum Data Burst Volume  (NOTE 2) | Default  Averaging Window | Example Services |
| 1 | GBR | 20 | 100 ms  (NOTE 11,  NOTE 13) | 10-2 | N/A | 2000 ms | Conversational Voice |
| 2 | (NOTE 1) | 40 | 150 ms  (NOTE 11,  NOTE 13) | 10-3 | N/A | 2000 ms | Conversational Video (Live Streaming) |
| 3 |  | 30 | 50 ms  (NOTE 11,  NOTE 13) | 10-3 | N/A | 2000 ms | Real Time Gaming, V2X messages (see TS 23.287 [121]).  Electricity distribution – medium voltage, Process automation monitoring |
| 4 |  | 50 | 300 ms  (NOTE 11,  NOTE 13) | 10-6 | N/A | 2000 ms | Non-Conversational Video (Buffered Streaming) |
| 65  (NOTE 9,  NOTE 12) |  | 7 | 75 ms  (NOTE 7, NOTE 8) | 10-2 | N/A | 2000 ms | Mission Critical user plane Push To Talk voice (e.g., MCPTT) |
| 66  (NOTE 12) |  | 20 | 100 ms  (NOTE 10,  NOTE 13) | 10-2 | N/A | 2000 ms | Non-Mission-Critical user plane Push To Talk voice |
| 67  (NOTE 12) |  | 15 | 100 ms  (NOTE 10,  NOTE 13) | 10-3 | N/A | 2000 ms | Mission Critical Video user plane |
| 75  (NOTE 14) |  |  |  |  |  |  |  |
| 71 |  | 56 | 150 ms (NOTE 11, NOTE 13, NOTE 15) | 10-6 | N/A | 2000 ms | "Live" Uplink Streaming (e.g. TS 26.238 [76]) |
| 72 |  | 56 | 300 ms (NOTE 11, NOTE 13, NOTE 15) | 10-4 | N/A | 2000 ms | "Live" Uplink Streaming (e.g. TS 26.238 [76]) |
| 73 |  | 56 | 300 ms (NOTE 11, NOTE 13, NOTE 15) | 10-8 | N/A | 2000 ms | "Live" Uplink Streaming (e.g. TS 26.238 [76]) |
| 74 |  | 56 | 500 ms (NOTE 11, NOTE 15) | 10-8 | N/A | 2000 ms | "Live" Uplink Streaming (e.g. TS 26.238 [76]) |
| 76 |  | 56 | 500 ms (NOTE 11, NOTE 13, NOTE 15) | 10-4 | N/A | 2000 ms | "Live" Uplink Streaming (e.g. TS 26.238 [76]) |
| 5 | Non-GBR | 10 | 100 ms  NOTE 10,  NOTE 13) | 10-6 | N/A | N/A | IMS Signalling |
| 6 | (NOTE 1) | 60 | 300 ms  (NOTE 10,  NOTE 13) | 10-6 | N/A | N/A | Video (Buffered Streaming) TCP-based (e.g., www, e-mail, chat, ftp, p2p file sharing, progressive video, etc.) |
| 7 |  | 70 | 100 ms  (NOTE 10,  NOTE 13) | 10-3 | N/A | N/A | Voice, Video (Live Streaming) Interactive Gaming |
| 8 |  | 80 | 300 ms  (NOTE 13) | 10-6 | N/A | N/A | Video (Buffered Streaming) TCP-based (e.g., www, e-mail, chat, ftp, p2p file sharing, progressive |
| 9 |  | 90 |  |  |  |  | video, etc.) |
| 69  (NOTE 9, NOTE 12) |  | 5 | 60 ms  (NOTE 7, NOTE 8) | 10-6 | N/A | N/A | Mission Critical delay sensitive signalling (e.g., MC-PTT signalling) |
| 70  (NOTE 12) |  | 55 | 200 ms  (NOTE 7,  NOTE 10) | 10-6 | N/A | N/A | Mission Critical Data (e.g. example services are the same as 5QI 6/8/9) |
| 79 |  | 65 | 50 ms  (NOTE 10,  NOTE 13) | 10-2 | N/A | N/A | V2X messages (see TS 23.287 [121]) |
| 80 |  | 68 | 10 ms  (NOTE 5,  NOTE 10) | 10-6 | N/A | N/A | Low Latency eMBB applications Augmented Reality |
| 82 | Delay Critical GBR | 19 | 10 ms (NOTE 4) | 10-4 | 255 bytes | 2000 ms | Discrete Automation (see TS 22.261 [2]) |
| 83 |  | 22 | 10 ms (NOTE 4) | 10-4 | 1354 bytes  (NOTE 3) | 2000 ms | Discrete Automation (see TS 22.261 [2]);  V2X messages (UE - RSU Platooning, Advanced Driving: Cooperative Lane Change with low LoA. See TS 22.186 [111], TS 23.287 [121]) |
| 84 |  | 24 | 30 ms  (NOTE 6) | 10-5 | 1354 bytes  (NOTE 3) | 2000 ms | Intelligent transport systems (see TS 22.261 [2]) |
| 85 |  | 21 | 5 ms  (NOTE 5) | 10-5 | 255 bytes | 2000 ms | Electricity Distribution- high voltage (see TS 22.261 [2]).  V2X messages (Remote Driving. See TS 22.186 [111], NOTE 16, see TS 23.287 [121]) |
| 86 |  | 18 | 5 ms  (NOTE 5) | 10-4 | 1354 bytes | 2000 ms | V2X messages (Advanced Driving: Collision Avoidance, Platooning with high LoA. See TS 22.186 [111], TS 23.287 [121]) |
| NOTE 1: A packet which is delayed more than PDB is not counted as lost, thus not included in the PER.  NOTE 2: It is required that default MDBV is supported by a PLMN supporting the related 5QIs.  NOTE 3: The Maximum Transfer Unit (MTU) size considerations in clause 9.3 and Annex C of TS 23.060 [56] are also applicable. IP fragmentation may have impacts to CN PDB, and details are provided in clause 5.6.10.  NOTE 4: A static value for the CN PDB of 1 ms for the delay between a UPF terminating N6 and a 5G-AN should be subtracted from a given PDB to derive the packet delay budget that applies to the radio interface. When a dynamic CN PDB is used, see clause 5.7.3.4.  NOTE 5: A static value for the CN PDB of 2 ms for the delay between a UPF terminating N6 and a 5G-AN should be subtracted from a given PDB to derive the packet delay budget that applies to the radio interface. When a dynamic CN PDB is used, see clause 5.7.3.4.  NOTE 6: A static value for the CN PDB of 5 ms for the delay between a UPF terminating N6 and a 5G-AN should be subtracted from a given PDB to derive the packet delay budget that applies to the radio interface. When a dynamic CN PDB is used, see clause 5.7.3.4.  NOTE 7: For Mission Critical services, it may be assumed that the UPF terminating N6 is located "close" to the 5G\_AN (roughly 10 ms) and is not normally used in a long distance, home routed roaming situation. Hence a static value for the CN PDBof 10 ms for the delay between a UPF terminating N6 and a 5G\_AN should be subtracted from this PDB to derive the packet delay budget that applies to the radio interface.  NOTE 8: In both RRC Idle and RRC Connected mode, the PDB requirement for these 5QIs can be relaxed (but not to a value greater than 320 ms) for the first packet(s) in a downlink data or signalling burst in order to permit reasonable battery saving (DRX) techniques.  NOTE 9: It is expected that 5QI-65 and 5QI-69 are used together to provide Mission Critical Push to Talk service (e.g., 5QI-5 is not used for signalling). It is expected that the amount of traffic per UE will be similar or less compared to the IMS signalling.  NOTE 10: In both RRC Idle and RRC Connected mode, the PDB requirement for these 5QIs can be relaxed for the first packet(s) in a downlink data or signalling burst in order to permit battery saving (DRX) techniques.  NOTE 11: In RRC Idle mode, the PDB requirement for these 5QIs can be relaxed for the first packet(s) in a downlink data or signalling burst in order to permit battery saving (DRX) techniques.  NOTE 12: This 5QI value can only be assigned upon request from the network side. The UE and any application running on the UE is not allowed to request this 5QI value.  NOTE 13: A static value for the CN PDB of 20 ms for the delay between a UPF terminating N6 and a 5G-AN should be subtracted from a given PDB to derive the packet delay budget that applies to the radio interface.  NOTE 14: This 5QI is not supported in this Release of the specification as it is only used for transmission of V2X messages over MBMS bearers as defined in TS 23.285 [72] but the value is reserved for future use.  NOTE 15: For "live" uplink streaming (see TS 26.238 [76]), guidelines for PDB values of the different 5QIs correspond to the latency configurations defined in TR 26.939 [77]. In order to support higher latency reliable streaming services (above 500ms PDB), if different PDB and PER combinations are needed these configurations will have to use non-standardised 5QIs.  NOTE 16: These services are expected to need much larger MDBV values to be signalled to the RAN. Support for such larger MDBV values with low latency and high reliability is likely to require a suitable RAN configuration, for which, the simulation scenarios in TR 38.824 [112] may contain some guidance. | | | | | | | |