**3GPP SA4 #129-e S4-241589**

**Online, 19 August 2024**

# Agenda item 10.6

# [FS\_5G\_RTP\_Ph2] On the feasibility of Application-layer FEC Awareness at the RAN

#### Qualcomm Incroporated

In the LS reply R2-2405781 [1], RAN2 states that “RAN2 does not think it is possible for NG-RAN to determine reliably whether a PDU has been successfully delivered over an unacknowledged-mode data bearer.”

The LS reply does not mean that AL-FEC awareness at the NG-RAN is infeasible because

* for MDS AL-FEC, the reliability of determining an individual PDU delivery status is sufficient, but not necessary for AL-FEC RAN awareness, and
* it is only concerned about the unacknowledged-mode data bearer, and there is also the acknowledged-mode data bearer that could be used.

In what follows, we elaborate on these two reasons.

For RAN awareness of AL-FEC, the AL-FEC codes under consideration have been MDS (or near MDS) codes. For MDS AL-FEC, the NG-RAN only needs to know how many PDUs, not which of this many PUDs, have been successfully delivered.

To determine how many PDUs have been successfully delivered by the NG-RAN, the NG-RAN needs to track the mapping of PDUs of a PDU Set to transport blocks (TBs). This tracking is not a new problem, because it is needed for the PDU Set based QoS as well.

**Observation 1:** For MDS AL-FEC codes, to drop obsolete PDUs within a PDU Set while allowing the UE to reconstruct all the source packets, it is the successful delivery of as many PDUs as the source packets and not the successful delivery of individual PDUs that the NG-RAN needs to determine reliabibely.

Furthermore, the reliability of the determination of the successful delivery of as many PDUs as the source packets can be increased if the NG-RAN collects more ACKs than necessary. Specifically, let there be K source PDUs. The mapping of PDUs of a PDU Set to the TBs may result in multiple PDUs being mapped to a same TB, and the HARQ feedback is on a per TB basis. Therefore, the gNB may receive X ACKs that acknowledge the delivery of K or more PDUs, where X ≤K. The mapping may also result in a PDU being segmented and put into multiple TBs. But this is not likely due to the increased latency and the low latency requirement of XR applications. The gNB, instead of receiving X ACKs, can receive X+M ACKs before deciding to stop transmitting other PDUs, where M≥1. It is explained next why the reliability can be increased. For ease of exposition, we assume each PDU is mapped to a separate TB and thus has its own HARQ feedback. The condition “or if impossible, all PDUs are sent” below is because it is possible that the gNB cannot receive any more ACKs after it has received X ACKs even if the gNB sends all PDUs of the PDU Set.

P(K or more PDUs are delivered successfully | K+M ACKs are received, or if impossible, all PDUs are sent) > P(K or more ACKs are delivered successfully | K ACKs are received, or if impossible, all PDUs are sent),

which is equivalent to

P(K or more ACKs are sent | K+M ACKs are received, or if impossible, all PDUs are sent) > P(K or more ACKs are sent | K ACKs are received, or if impossible, all PDUs are sent),

which is true because the event “K or more PDUs are delivered successfully” can be considered essentially the same as the event “K or more PDUs are sent” given that the receiver can determine whether a TB is delivered successfully or not reliably with the long CRC (16 or 24 bits).

To illustrate, we show the effect of M on the AL-FEC decoding and the savings on the data transmission for the cases of (1) K=10, N=15, (2) K=20, N=30, and (3) K=30, N=45, where N is the total number of PDUs after AL-FEC encoding of a PDU Set.

For the evaluation, the unreliability of individual HARQ feedback is characterized by two probabilities: NACK to ACK error, and ACK to NACK error. These error probabilities along with the respective maximum allowed values are defined in 3GPP TS38.104 and shown in Table x.

Table x HARQ feedback errors and the maximum allowed error probabilities

|  |  |  |
| --- | --- | --- |
| HARQ feedback errors | Definition | Maximum allowed error probability |
| NACK to ACK error | A NACK bit sent by the UE is mis-interpreted as an ACK bit by the NG-RAN | 0.1% |
| ACK to NACK error | An ACK bit sent by the UE is mis-interpreted as a NACK bit by the NG-RAN | 1% |

As shown in Figure 5.4.2.x-1:

* Subplot (a) shows that when the base station collects exactly K ACKs, the probability of AL-FEC decoding failure is below 0.4% for all cases, between 0.1% and 1% in Table x.
* When the base station collects just one extra ACK, the probability of AL-FEC decoding failure is reduced, and the reduction is more than one order of magnitude for K=20 and 30, with both probabilities of failure below 0.01%.
* As M increases, the probability of AL-FEC decoding failure quickly flattens out.
* Subplot (b) shows that the enhanced AL-FEC decoding performance is at the expense of transmitting more PDUs by the base station.

|  |  |
| --- | --- |
| (a) | (b) |

Figure 5.4.2.x-1: For the RLC UM mode, the impact of M (the number of extra ACKs the base station collects) on (a) the probability of AL-FEC decoding failure, and (b) the average number of intentionally dropped obsolete packets.

**Observation 2:** For RLC UM and MDS AL-FEC codes, NG-RAN can increase the reliability of determining whether as many PDUs as the source PDUs are successfully delivered by receiving more ACKs than necessary. Specifically, the NG-RAN receives X+1 or more ACKs, where X is the number of ACKs that acknowledge K or more PDUs while X-1 ACKs do not, and K is the number of srouce packets.

On the RLC mode in the RAN2 LS reply, it refers to the UM mode. It does not rule out the possibility of RLC AM configured for low latency. In fact, RAN is currently working on improving RLC AM to serve low-latency traffic under RAN\_XR\_Ph3 [2]:

*Specify the following user plane enhancements [RAN2]*

*- RLC re-transmission related enhancements for operation of RLC Acknowledged Mode (AM) with small packet delay budget.*

**Observation 3:** The RAN2 LS R2-2405781 does not rule out the possibility that RLC AM can be used for low-latency traffic.

To summarize, we have:

**Observation 4:** The SA4 understanding of the RAN2 LS reply in R2-2405781 is as follows:

1. For RLC UM and MDS codes, there is no need to have 100% reliability of invidivual PDU delivery, as long as the probability that as many PDUs as the source PDUs are successfully delivered is high enough.
	* The probability can be improved if the NG-RAN collects redundant ACKs.
2. The way to achieve low latency for XR traffic is not necesaryily through RLC UM, and the RAN2 LS R2-2405781 does not rule out RLC AM.

**Proposal:** SA4 sends an LS to SA2 to confirm point 1 in Observation 4.

**References**

[1] R2-2405781, "Reply LS on Application-Layer FEC Awareness at RAN", 3GPP SA WG2 Meeting #126, Fukuoka, Japan, 20-24 May 2024.

[2] RP-240791, "Revised WID on XR (eXtended Reality) for NR Phase 3", 3GPP RAN Meeting #103, Maastricht, Netherlands, March 18-21, 2024.