3GPP TSG RAN WG1 Meeting #93 **R1-1808256**

Gothenburg, Sweden, August 20th – 24th, 2018

Agenda Item: 7.2.6.4
Source: **TCL Communication**
Title: **Prediction-Based early feeback**
Document for: **Discussion and decision**

1. Introduction

The requirements of URLLC have been specified in [1]:

"A general URLLC reliability requirement for one transmission of a packet is 10^{-5} for 32 bytes with a user plane latency of 1ms."

In last RAN1 meeting, the following agreements about PDCCHs and PDSCH were made [2]:

Agreements:

- *To ensure the reliability requirement of NR-PDCCH for URLLC, at least the following aspects should be supported*
  - *Defining a compact DCI format targeting low BLER operation*
  - *The highest aggregation level should target a BLER of Y for this compact DCI format*
    - *FFS Y, Y<1%*
    - *FFS highest aggregation levels, e.g., 16, 32*
  - *FFS other enhancements*

[3] gives the topics that will be studied in the next release and one item about control channel design is:

- *Study and specify if gains are identified*
  - *Define a new DCI format(s) that has a smaller DCI payload size than DCI format 0-0 and DCI format 1-0 unicast data*
  - *For a given carrier, PDCCH repetitions over same or multiple PDCCH monitoring occasion(s) of the same or multiple CORESET and search space*

2. Early prediction-based PDSCH feedback

HARQ process allows the gNB to carry out the retransmissions to increase the overall reliability of the transmission in the specified time constraint when the codewords in the initial transmission and the retransmissions are combined to generate a new codeword with better SNR and/or a lower code rate. However, the number of retransmissions is limited because of HARQ latency and latency requirement of URLLC. Thereby, HARQ latency must be reduced to create more retransmission occasions. HARQ latency consists of:

- τ: propagation delay
- T_{TTI}: transmission time interval duration
- T_{FB}: generating feedback time that includes the decoding time for the whole received signal
- $T_{A/N}$: transmission time for ACK/NAK
- T_{Tx}: processing time of the feedback at the gNB

In these terms, T_{FB} can be reduced to lower HARQ latency. The other terms are fixed due to the inherent characteristics or hard to reduce. τ depends on the distance between the gNB and the UE that cannot be changed. Similarly, $T_{A/N}$ and T_{TTI} are also fixed. T_{Tx} is hard to improve when normally feedback contains only 1 bit so the decoder at the gNB is already able to decode HARQ feedback quickly.

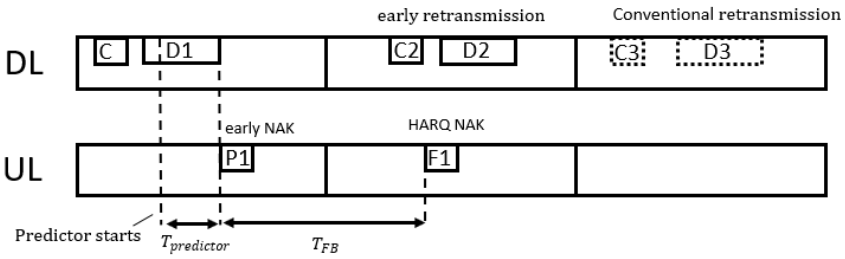

Figure 1: Prediction-based feedback and regular HARQ feedback

As can be seen in Figure 1, T_{FB} has heavy contributions from the transmission time during which the transmitter is transmitting the packet and the processing time at the receiver during which the receiver is doing the receive processing. The receive processing includes but not limited to equalization, demodulation and channel decoding where channel decoding at the decoder is quite an onerous task. It impedes the opportunity of the useful retransmissions in the time constraint.

In an effort to optimize the HARQ feedback timing and creating more (re-)transmission opportunities within a certain latency target, we propose to use a two-stage feedback. The first stage of feedback is basically a prediction about the success/failure of the transport block. This early prediction-based feedback is designed so that it can be transmitted by the receiver very quickly, possibly even before the complete reception of the transport block of data. It informs the transmitter about the result of the channel decoder by an intelligent estimation without passing through the whole decoding process. The predictor evaluates the error probability based on LLR estimation by using a fraction of the transmitted transport block instead of using the message passing algorithm for the whole codeword as in the decoder. It reduces computation complexity of the predictor and makes decision-making time decrease significantly compared to the full decoding process. The second stage of feedback is a conventional HARQ feedback.

To reduce the impact of the TTI duration in the HARQ RTT, the scheme proposes that the receiver only uses a fraction of the transport block signal to predict the outcome. This means that the receiver does not need to wait until the end of TTI for the complete reception of the transport block to start process (decode) the codeword, rather it can start the prediction computation just after receiving a fraction of the transport block, way before the complete reception of the transport block. In HARQ process, time from the arrival of data in the UE to the generation of feedback is calculated by:

$$T = T_{TTI} + T_{FB} \quad (1)$$

On the other hand, in early feedback process, time from the beginning of data transmission in the gNB to the generation of feedback in the UE is calculated by:

$$T' = r_p \times T_{TTI} + T_{predictor} \quad (2)$$

where r_p: the ratio between the code used for prediction and the whole transmitted code

(1) and (2) show that T' is much smaller than T when only a fraction of T_{TTI} is necessary for prediction. Another reason is that $T_{predictor}$ is also smaller than T_{FB} when the computation in the predictor is less complex than that in the decoder. Thus, early feedback can be generated before the UE receives the whole transmitted signal from the gNB and the retransmission is likely to start much earlier than the scheme with regular HARQ feedback.

Proposal 1: The UE uses a part of the received signal to estimate the error probability of the decoding process based on LLR estimation and generates early feedback to the gNB. The gNB can use early feedback to trigger an immediate retransmission.

In another method, to further improve the reliability of the prediction, the user can combine the earlier received signal as well in addition to the fraction of the coming transport block for which prediction feedback is being prepared. As a concrete example in DCI based DL transmission, if the user receives DCI which schedules the data for which user has to send the two-stage feedback. For the first stage prediction-

based feedback preparation, the user will use the fraction of the transport block signal. To improve the reliability of the prediction-based feedback, the user can combine this signal with the DCI as well. As an example, it can use the SNR of its detected DCI or some other metrics improve the probability of the correct prediction.

Proposal 2: The UE can use other metrics from earlier signal as decoded DCI to support the reliability of the prediction.

Early prediction feedback can be very helpful to indicate the gNB not only about potential PDSCH failure but also PDCCH failure. As shown in , the early prediction-based feedback will make the transmission much more robust against the PDCCH errors as an absence of prediction feedback at the gNB will be a direct indication of missed PDCCH.

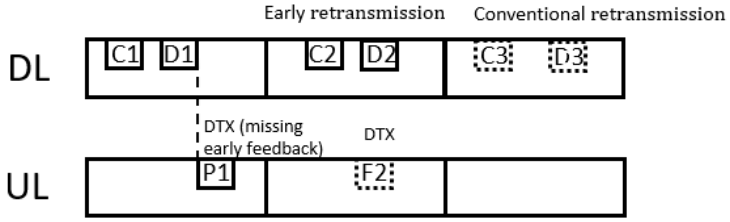

Figure 2: Missing early feedback triggers an early transmission

Proposal 3: Early prediction feedback is used as an indicator of a success or a failure in decoding PDCCH. In case an early feedback is missing, the gNB detects a DTX and starts a retransmission immediately instead of waiting for a normal time interval of HARQ feedback.

The prediction allows the receiver to transmit a very fast response to the transmitter regarding the success/failure of the transport block, even before receiving the full data of the transport block. This helps the gNB have more chances to retransmit the packet to boost the reliability. However, the predictor also makes the system suffer from false prediction. There are two kinds of false prediction: false negative (FN) and false positive (FP). False negative occurs when an early NAK is sent while the decoder decodes correctly the codeword. It causes a waste of resources due to the unnecessary retransmissions but it does not affect the reliability directly. False positive occurs when an early prediction ACK is sent but, in fact, the decoder fails to decode data. This means that there is no retransmission and the packet is lost. It affects the performance of URLLC transmission. Therefore, false positive is more severe than false negative. The probabilities of false negative and false positive can be adapted by changing the proper threshold in predicting ACK or NAK. The ratio between false positive and false negative is changed depending on requirements and tolerance of the system.

Proposal 4: Threshold in feedback prediction is adapted to control the probabilities of false negative and false positive.

3. Transmission schemes to alleviate the effect of false prediction

In order to avoid the harmful effects of false prediction and take advantage of early feedback's benefits, a scheme combining both early feedback and regular HARQ feedback is proposed. The gNB is able to use HARQ feedback and then switches to early feedback when a fast retransmission is required to achieve the reliability requirement. The moment to switch from single stage classic HARQ feedback to two-stage feedback consisting of the first stage of early prediction feedback and the second stage of classic feedback is decided by the gNB. When two-stage feedback is activated, the gNB has to inform the users about the resources for both stages of feedback. To keep the things simple, it would make sense to consider early prediction-based feedback similar to legacy feedback for encoding purpose at the user. Therefore, the user

can apply the same encoding and transmit processing for the early prediction-based feedback as of the classical HARQ feedback. This activation can be sent in the higher layer signaling to the user.

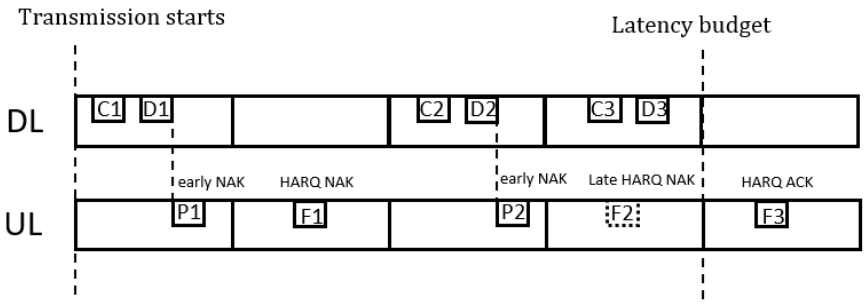

Figure 3: Downlink transmission with early feedback and the gNB sensing latency budget

As can be seen in Figure 3, in the first transmission, the gNB sends PDCCH and PDSCH (C1 and D1, respectively) then the UE predicts a failure of the decoder and transmits an early NAK (P1) but the gNB still waits HARQ feedback (F1) to confirm that failure and retransmits the packet (C2 and D2) because it senses the remaining latency budget and recognizes that it still has enough time left to reach the target reliability with the conventional classical HARQ feedback. In the retransmission, the UE continues to predict a failure of the decoder (P2). This time, the gNB senses that latency budget is not left much and a useful retransmission is impossible in the time constraint if it waits classical HARQ feedback (F2). For this reason, in case data is actually not decoded correctly, the packet will be lost. Therefore, the gNB reacts very fast to this early feedback to trigger an immediate retransmission (C3 and D3) to increase the chance that the UE can decode data correctly and the reliability of the system is boosted.

Proposal 5: The gNB senses latency budget following URLLC requirement so as to decide to use early feedback or HARQ feedback. The gNB only uses early feedback when latency budget does not remain enough to wait HARQ feedback in order to make decision about a retransmission.

An alternative scheme is also considered when it causes the waste of resources but creates more retransmission occasion than the above scheme. If the gNB receives early ACK, it does not take that feedback into account and continues to wait for HARQ feedback in order to decide to terminate or retransmit data. Therefore, the system avoids suffering from losing packet due to false positive. On the other hand, if the gNB receives early NAK (P1) as in Figure 4, it carries out an immediate retransmission (C2 and D2). After that, if HARQ feedback is NAK, that retransmission still continues. This means that the early retransmission can be translate to more transmission occasions if data continues not to be decoded correctly. In contrast, if HARQ feedback is ACK (F1), that retransmission is no longer necessary. As illustrated in Figure 4, the gNB will stop that retransmission (C2 and D2) instantly after receiving ACK HARQ feedback to prevent from wasting resources and leaves resources for other UEs.

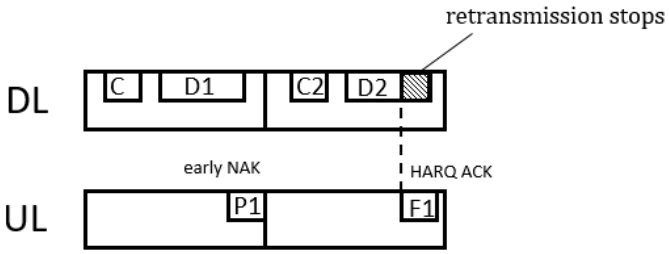

Figure 4: The gNB stops a retransmission triggered by early NAK after receiving ACK HARQ feedback

Proposal 6: The gNB only takes early NAK into account. The retransmission is carried out immediately after receiving an early NAK. This retransmission continues if the gNB receives NAK HARQ feedback later. On the other hand, if the gNB receives ACK HARQ feedback that means a false negative of early feedback, the gNB stops immediately the retransmission to reduce to waste resource for the unnecessary retransmission and the transmission completes.

4. Simulation result

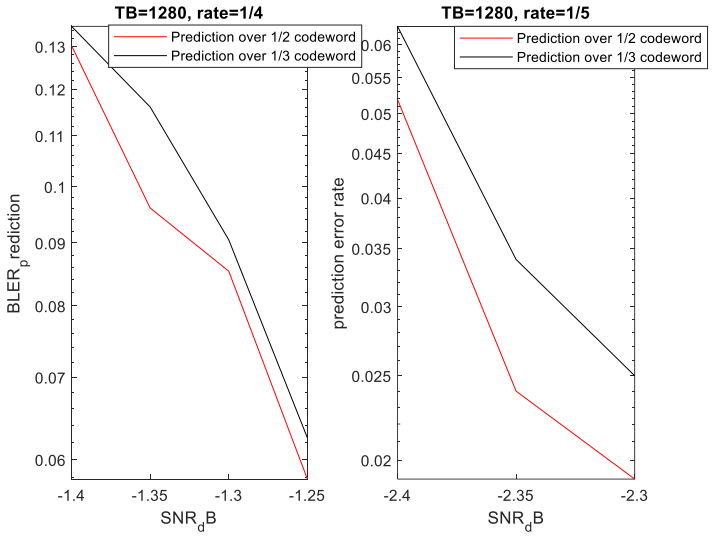

Figure 5: Prediction error rate

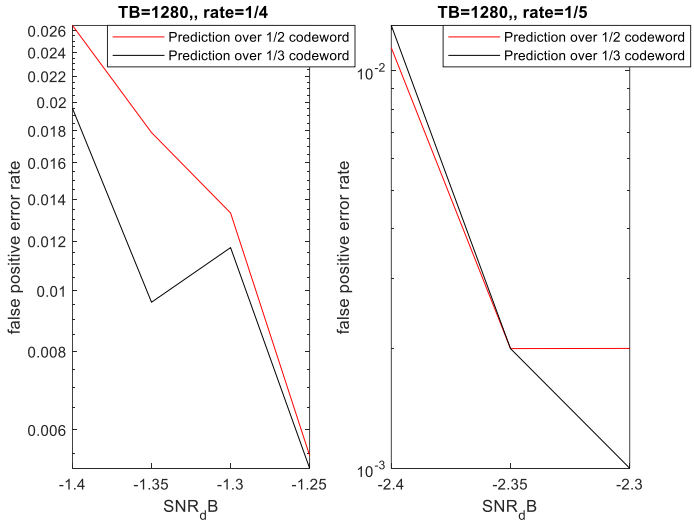

Figure 6: Error rate's false positive

The simulation of prediction error rate is shown in Figure 5. In this simulation, an input codeword with size 1280 is encoded by base graph 2 of LDPC code as agreed in 3GPP standard with rate 1/4 and 1/5. The encoded codeword is modulated by QPSK and transmitted in AWGN channel. The decoder uses message passing algorithm with min-sum calculation to decode the incoming codeword. The maximum iteration of the decoder is 25. The predictor also uses the same algorithm as the decoder but with fewer iterations that is 5. In the predictor, two cases are considered where a half and a third of the transmitted codeword is used

to estimate the outcome, respectively. The prediction error including both false negative and false positive is calculated when block error rate (BLER) of the whole codeword is approximate to 10^{-2}. The predictor works much better with a lower rate. The reason is that at a lower rate, the codeword is longer so the portion of the codeword that is used to estimate the error probability is also longer. The predictor has more information and the sub-codeword in the predictor has a higher probability to converge.
For the same simulation, but the error rate of false positive is considered separately in Figure 6. As shown in the graph, the error rate of false positive is much smaller (around 10 times smaller) than the overall prediction error. A small false positive has very positive impact in improving the QoS requirements for the URLLC. As an example, the above figure shows that for a transmission with target BLER of 10^{-2}, the false positive prediction error of 10-3 is achievable which means that in the error cases, the early prediction would already have requested the retransmission from the gNB and only in 0.1% of the cases, the gNB would have to trigger the retransmissions after the 2nd stage classical feedback. This certainly comes at the price of the increased resource utilization of the two-stage feedback but for critical URLLC traffic when there are requirements to meet a certain reliability within a certain latency, this can provide very useful means to achieve such targets. Moreover, two proposed schemes in part 3 also reduce the effect of false negative when the gNB decides to use NAK prediction by sensing latency budget or stops the retransmission after receiving ACK.

5. Conclusion

Proposal 1: The UE uses a part of the received signal to estimate the error probability of the decoding process based on LLR estimation and generates early feedback to the gNB. The gNB can use early feedback to trigger an immediate retransmission.
Proposal 2: The UE can use other metrics from earlier signal as decoded DCI to support the reliability of the prediction.
Proposal 3: Early prediction feedback is used as an indicator of a success or a failure in decoding PDCCH. In case an early feedback is missing, the gNB detects a DTX and starts a retransmission immediately instead of waiting for a normal time interval of HARQ feedback.
Proposal 4: Threshold in feedback prediction is adapted to control the probabilities of false negative and false positive.
Proposal 5: The gNB senses latency budget following URLLC requirement so as to decide to use early feedback or HARQ feedback. The gNB only uses early feedback when latency budget does not remain enough to wait HARQ feedback in order to make decision about a retransmission.

Proposal 6: The gNB only takes early NAK into account. The retransmission is carried out immediately after receiving an early NAK. This retransmission continues if the gNB receives NAK HARQ feedback later. On the other hand, if the gNB receives ACK HARQ feedback that means a false negative of early feedback, the gNB stops immediately the retransmission to reduce to waste resource for the unnecessary retransmission and the transmission completes.

References

[1] "TR 38.913".

[2] "Chairman's Notes RAN1 NR Ad-Hoc #1".

[3] "RP-172817 – “NR High-Reliability URLLC scope for RAN1/RAN2”, Ericsson, RAN78".