RAN-R18-WS-crossFunc-Futurewei - Version 0.0.2
RAN

3GPP TSG RAN Rel-18 workshop RWS-210621

Electronic Meeting, June 28 - July 2, 2021

Agenda Item: 4.3

Source: Futurewei

Title: Email discussion summary for [RAN-R18-WS-crossFunc-Futurewei]

1 Introduction

This email discussion summary covers the comments and questions and answers on the following Futurewei contributions:

RWS-210038 ML/AI for wireless;

RWS-210040 SDAP Enhancement for Low Latency High Throughput Communication.

2 RWS-210038 ML/AI for wireless

Even though AI/ML shows promising results in many areas, its benefits and justification haven't been thoroughly investigated for PHY layer problems. RWS-210038 proposes a Rel-18 SI to identify candidate AI/ML-based PHY layer use cases, study standards impact and evaluation methodology for performance and robustness and build common datasets to facilitate the study and evaluation.

2.1 Round 1 Comments on RWS-210038

Feedback Form 1: General Comments

1 – Rakuten Mobile We support most of the proposals in RWS-210038 and also few more like categorization of AI enhancements based on co-ordination requirements and Interoperability.
2 – Fujitsu Limited Thanks for the contribution, we have several questions for clarifications: 1. In your CSI feedback compression simulation, is there performance comparison between compressed CSI vs. Type1 or Type2 codebook. 2. What is your view on how to select use case for AI? 3. What is your view on how to set common datasets? for example how to make dataset for training and dataset for inference?

3 – Nokia Corporation

Thank you for the contribution, we have some questions for clarifications:

Q1: Given your note that learning/inference may be done at UE and network, do you propose to first study federated/distributed learning in RAN3 or also as part of RAN1/2 air interface studies?

Q2: Regarding beam management applications, how do you define the optimal beam?

Q3: Where do you propose to train the optimum beam prediction model, at UE or network?

Q4: Similarly, for mmWave blockage, where do you propose to train the prediction model, at UE or network?

4 – HUAWEI TECHNOLOGIES Co. Ltd.

Thank you very much for the contribution. In general we think sufficient study on aspects like evaluation methodology, input/output for model training and model inference etc should be done first before we can conclude whether/which use case is beneficial with AL/ML. Here we have some question for clarification:

Q1. In Page 4, the output of encoder are binary bits or real number?

Q2. In Page 7, what will be feedback from UE to gNB?

5 – Apple GmbH

We in general agree with the proposals. However, we would also want to know what the Ue will feedback to the network. Additionally, do you have any details on how to resolve conflicts with the AI/ML engines running on both the network and the UE sides.

2.2 Round 1 Questions on RWS-210038

Feedback Form 2: Round 1 Questions on RWS-210038

1 – Xiaomi Communications

[Xiaomi] Thank you for the contribution. We think the joint optimization of multi-functions is interesting but may be more chanllenging . What is your consideration on how to coordinate the study of individual block optimization and the study of multi-function optimization ?

2 – ZTE Corporation

One question on “Multi-functional-block PHY optimization”: which blocks in the PHY chain can be jointly optimized with AI in your view?

3 – Sony Corporation

Thanks for the contribution. We have two questions.

- AI neural encoder and decoder
  - How are the AI neural encoder and decoder trained. For example, network provides some data set for training, or trained AI model is downloaded to the UE?

- Joint optimization of multi-functions at PHY
  - What PHY functions should be prioritised to be jointly optimised?

4 – CAICT

Thanks for the good contribution. Establishing common test datasets is a good idea and could you please provide some further explaination? For example, the data type and supporting use cases for the common test datasets.

5 – MediaTek Inc.

Thanks for the contribution. some questions for clarification:

1) Does Futurewei have any priority among the use cases mentioned to star with?

2) What are the tradeoffs mentioned in slide 12 expected by Futurewei?

3) Does Futurewei expect a common data set specific to applications or will it be cross multiple applications?

6 – NTT DOCOMO INC.

Thank you for the contribution. We are interested in AI-based link adaptation enhancement.

To get us familiar, we have some questions about AI-based MCS selection.

Is AI performed on BS in this scheme? Do you think this scheme requires additional signaling between UE and gNB?

2.3 Round 1 Answers on RWS-210038

To Rakuten Mobile's comment:

Thanks for your support. We can discuss further regarding what additional areas can be part of the Rel-18 SI.

To Fujitsu's questions:

Thanks for the questions. Followings are our responses:

[Rsp for Q1] We plan to compare AI/ML-based CSI compression and reconstruction performance with a few baselines and analytical based approaches.

[Rsp for Q2] In terms of selecting PHY layer use cases, our view is to start from those that are computationally complicated / challenging, or those complicated non-convex problems that are lacking optimal analytical solutions, i.e., only sub-optimal solutions exist first.

[Rsp for Q3] As AI/ML-based approaches are data-driven, we believe common datasets will facilitate performance evaluation and baseline and benchmark comparisons. For "some" selected use cases, we suggest companies can jointly contribute and build these common datasets together.

For each identified use case, input and output to/from AI/ML functional modules (training and inference

functional blocks) have to be clearly defined. Once input/output are specified, each company who would like to evaluate performance (based on some agreed-upon methodologies/critera) of their approaches may choose to use one (or more) of the datasets and share their results if desired.

We believe estblishing some common datasets will greatly help the evaluation and benchmark comparison process for AI/ML-based approaches.

To Nokia's questions:

Thanks for the questions. Followings are our responses:

[Rsp for Q1] We would like to get some clarification from Nokia regarding this question whether it is for a particular PHY layer use case or this is a general question.

[Rsp for Q2] For AI/ML-based beam selection, we think the answer depends on how each company would like to formulate it into a learning problem. Different learning approaches may choose to use different mechanisms to determine or approximate the optimal beam.

[Rsp for Q3] Regarding whether the leraning for beam selection should be performed on the network side or UE side, it depends on the use case (i.e., it can be on the UE side or on the network side). If the use case is for the gNB to determine the best beam (or a set of candidate beams initially), then the learning can be performed mainly on the network side.

[Rsp for Q4] For mmWave blockage problem, the solution may also be use case (direction) dependent. For the diagram and proposal described in our contribution, we propose to study the learning based approach at network side.

To Huawei's questions:

Thanks for the questions. Followings are our responses:

[Rsp for Q1] The format of the output for the encoder belongs to a design decision. For example, the output can be a vector with M real numbers, or the output can be a sequence of bits. We suggest to discuss the details once the scope of the SI and a set of candidate use cases are agreed among companies.

[Rsp for Q2] The diagram in our contribution for beam selection use case represents one design option while there can be other alternatives. The arrow from UE to gNB represents the signal received from UE at gNB after applying receive beamforming.

To Xiaomi's question:

Thanks for the question. We agree that multi-functional-block PHY optimization is more challenging/complicated. We believe there is no conflict between studying use cases for individual functional block optimization and multi-functional-block optimization. We think studying both cases will allow us to better determine what use case(s) are beneficial when considering short and long term objectives and taking into account the results after comparing with analytics-based approaches (both performance and computation complexity reduction).

To ZTE's question:

Thanks for the question. One example scenario is the end-to-end optimization. In this scenario, the information bits can be considered as input to a AI/ML "information encoder" (for multi-antenna) while the

receiver side has a ”information decoder” (also for mult-antenna), followed by a reconstruction function to recover the original information bits.

The example use case described above is an end-to-end optimization while there are other use cases (non end-to-end) which may jointly optimize some functional blocks together, and one example would be joint source and channel coding optimization.

To Sony’s questions:

Thanks for the questions. Followings are our responses:

[Rsp for Q1] We believe there are multiple ways to train the intial encoder+decoder or autoencoder. One option would be to collect some sample data for training then offload the encoder and decoder to proper entities in the network.

[Rsp for Q2] One example scenario is the end-to-end optimization. In this scenario, the information bits can be considered as input to a AI/ML ”information encoder” (for multi-antenna) while the receiver side has a ”information decoder” (also for mult-antenna), followed by a reconstruction function to recover the original information bits.

There are other use cases that involve optimization of multiple PHY functionalities, for example joint source and channel coding. At this pre-SI stage, we are open for your feedbacks.

To CAICT’s question:

Thanks very much for your comments. As AI/ML-based approaches are data-driven, we believe common datasets would greatly help the performance evaluation and benchmark comparison process for AI/ML-based approaches.

As each use case may have different input and output requirements (even for raw data features), common datasets can be estbalished per use case, or a few related use cases together. We can discuss about the format for the common dataset in more details once candidate use cases are agreeed. Providing raw attributes is one option, or an agreed-upon encoded format is another option. There can be other alternatives as well. At this stage, we are open for your feedbacks.

To MediaTek’s questions:

Thanks for the questions. Followings are our responses:

[Rsp for Q1] In terms of selecting PHY layer use cases, our view is to start from those that are computationally complicated / challenging ones as they require more computing power and time to solve, or those complicated non-convex problems that are lacking optimal analytical solutions, i.e., only sub-optimal solutions exist first.

For those use cases we proposed, we believe they are good candidates while we consider CSI compression/reconstruction and channel estimation are good starting point.

[Rsp for Q2] Typically there is no ideal solution that can achieve optimal performance while with very low overhead. This is no different from analytics based approach. For a given use case, each AI/ML-based approach may have its strenghth and weakness and we need to invetigate them to reach some balance. One general tradeoff is to balance between performance and air interface overhead, for example, if one AI/ML approach can achieve better performance but require a lot of data to be tranferred over the air interface while another approach achieves slightly worse performance but can reduce air interface overhead significantly, then

we may want to find a proper approach that can strike a balance based on some criteria.

[Rsp for Q3] As each use case may have different input and output requirements (even for raw data features), common datasets can be estbalished per use case, or a few related use cases together. Once candidate use cases are agreed among companies, we can discuss the details whether multiple use cases can share and use one common dataset. At this stage, we are open for your feedbacks.

To NTT DoCoMo's questions:

Thank you for being interested in our contribution. Please see our responses below.

[Rsp to Q1] In this proposal we are mainly proposing a problem to be solved, the link adaptation. Although we have suggested different ways of applying AI/ML to the problem, we are not proposing any specific scheme to solve the problem. But, yes, in general, we think AI will perform on the BS side.

[Rsp to Q2] It depends. As we proposed, there could be two types of solutions. A more evolutional one will target at using ML to enhance the performance of ILLA and/or OLLA. In this case, it may not need additional signaling; the other type of solution, which is more revolutional, will try to use ML to replace the ILLA and/or OLLA. In this case, additional signaling may be needed. But, overall, these are the topics to be studied if we decided to work on the topic. We invite NTT Docomo to work on the issue together if you have interest on the topic.

To Apple's questions:

Thanks for your support. Regarding your question, we would like to get some clarification:

1) For UE feedback, is there a specific use case you are referring to?

2) Could you please clarify what "conflicts" means in this context?

2.4 Round 2 Questions on RWS-210038

Feedback Form 3: Round 2 Questions on RWS-210038

1 – vivo Mobile Communication Co.

Thanks for sharing your thoughts.

We have some questions as below

Q1. Regarding the beam selection procedure in Slide7, can you elaborate a little bit on how the beam is selected? What kind of feedback do you envision for the last step in your illustrated procedure?

Q2: The paper focuses on PHY layer enhancement. For those air interface enhancement that is not directly related to PHY (e.g., for mobility related enhancement), what is your view?

Q3: For the study item arrangement, is it correct understanding that you prefer to do evaluation for all the use cases and the downselection is based on the evaluation? Do you think we need a phase to collect all potential use cases?

2 – Samsung Electronics Polska

1. In page 5 for the results for AI CSI compression, what is the channel for the training data? E.g. simulation channel modeling like TDL, CDL,38.901? Or measured data collected in practice?
2. In page 8, for mmWave blockage prediction.

The output of ML model is probability of blockage in future intervals. How long is the future intervals here? several ms, hundreds of ms or serveral seconds? Also, what is the pre-blockage characteristic?

2.5 Round 2 Answers on RWS-210038

To Vivo's questions:

Thanks for the questions. Followings are our responses:

[Rsp to Q1] As AI/ML-based approach is data-driven, we think the AI/ML-based beam selection algorithm may have multiple options in choosing the optimal beam, which could be design-dependent. For example, if the goal is to optimize channel capacity, then the algorithm may choose the beam that has the highest predicted channel capacity. The feedbacks can also be the measurements related to channel capacity or other measurements that can help estimate the channel capacity.

[Rsp to Q2] Our view in selecting candidate use cases for the SI can start from those that are computationally complicated / challenging, as they require more computing power and time to solve, or those complicated non-convex problems that are lacking optimal analytical solutions, i.e., only sub-optimal solutions exist. Some AI/ML-based mobility enhancement related functionalities have been covered under Rel-17 RAN3 SI "Enhancement for Data Collection for NR and EN-DC", which is a good starting point and we expect Rel-18 may start the normative work specified in this SI .

[Rsp to Q2] Our view in selecting candidate use cases for the SI can start from those that are computationally complicated / challenging ones as they require more computing power and time to solve, or those complicated non-convex problems that are lacking optimal analytical solutions, i.e., only sub-optimal solutions exist. Some AI/ML-based mobility enhancement related functionalities have been covered under Rel-17 RAN3 SI "Enhancement for Data Collection for NR and EN-DC" which is a good starting point and we expect Rel-18 may start the normative work specified in this SI .

[Rsp to Q3] We have explained the high level principles of selecting use cases in our response for the 2nd question. For this Rel-18 SI, many companies have proposed candidate use cases based on their experience and knowledge that those use cases are challenging for analytics-based approaches. Thus, we could investigate these candidates and select a subset of the use cases from this pool to be part of the Rel-18 SI.

To Samsung's questions:

Thanks for the questions. Followings are our responses:

[Rsp to Q1] The sample data we used was generated from simulation and TDL-A channel model was used.

[Rsp to Q2] The granularity and blockage time interval(s) in the prediction output depend on the granularity of the measurements and design of the algorithm, which may be adjusted if needed.

The pre-blockage characteristics are the patterns that the AI/ML model will learn from the measurements prior to the blockage in the training data. Once the AI/ML model is trained, future blockages can be predicted using

the newly collected measurements.

3 RWS-210040 SDAP Enhancement for Low Latency High Throughput Communication

SDAP (Service Data Adaptation Protocol) was introduced and specified in R15 for NR to provide 5G RAN with capabilities of QoS control through QoS flow to DRB mapping. In R16 and R17, except for SDAP extension to sidelink, there has been no improvement on SDAP's functionalities in RAN, while SA has been introducing more flow based QoS measurement and enforcement. In order to meet increasing need of latency critical high throughput applications, *it is proposed to* ***enhance SDAP in R18 with latency and data rate awareness***.

3.1 Round 1 Comments on RWS-210040

Feedback Form 4: General Comments

1 – Intel Corporation (UK) Ltd 1) Is "mapping one QoS flow to multiple DRBs" a pre-requisite of active queue management of DRBs? 2) Regarding "N:M mapping from QoS flow to DRBs", how to guarantee in order delivery in one QoS flow which is mapped to multiple DRBs given that reordering is performed per DRB?
2 – ROBERT BOSCH GmbH The TDoc mentioned survival time (which is part of TSC Assistance Information (TSCAI) in the QoS).
- Q1: Do you support special handling to TSC in SDAP? or you prefer only consider QoS Flow as it is?
- Q2: Do you support MAC to be aware about QoS Flow? similar to Q1, would you support TSC to separately identified to MAC (may be also MDBV separately, etc.)?
- Is "mapping one QoS flow to multiple DRBs" always needed? Should we allow one DRB not mapped to non-TSC traffic (if exist) at least by implementation ?
3 – Apple GmbH Our primary concern here is with relays in the middle. Can we still ensure this with relay node architectures?

3.2 Round 1 Questions on RWS-210040

Feedback Form 5: Round 1 Questions on RWS-210040

1 – Qualcomm CDMA Technologies
What is the specific L2 metric and how is it measured? Is the intention to allow such L2 metric to be visible to QoS flow?

3.3 Round 1 Answers on RWS-210040

To Intel's questions:

Thanks for the questions.

1) No, mapping a QoS flow to multiple DRBs is not a pre-requisite of active queue management. Even when a QoS flow is mapped to one DRB, active queue management can still be performed to make sure that the queue of the DRB doesn't exceed a configured threshold. If the queue is full, SDAP doesn't deliver packets of the QoS flow to the DRB.

2) When a QoS flow is mapped to multiple DRBs, the in-order delivery is maintained per DRB, not between packets delivered on different DRBs. This is in line with current internet technologies of using multiple streams to reach endpoints independently, e.g., HTTP/2's multiplexed connections, supported by Chrome, Microsoft Edge, Firefox, Safari, etc. It targets at applications requiring low latency and high throughput.

To Robert Bosch GmbH's questions:

Thanks for the questions.

1) TSC requirements are also defined per QoS flow. SDAP entity operates on QoS flow, and should be able to take TSC requirement into account.

2) As discussed in slide 5 of RWS-210040, unless configuring DRB per QoS flow, MAC doesn't have visibility of a QoS flow and its traffic burst. For example, if 2 or more QoS flows are mapped onto one DRB, MAC would not be able to measure MDBV of individual QoS flows.

3) No, a QoS flow doesn't have to be mapped to multiple DRBs; it can be mapped to one DRB, as already supported since Rel-15. If and how a QoS flow is mapped to one or more DRBs is up to network configuration. And network can choose to map TSC and non-TSC traffic onto separate DRBs. However, there is a maximum number of DRBs a UE is required to support, given the processing and memory resource required when a DRB is configured. It may not be practical to dedicate one DRB for each QoS flow, when UE needs to support a number of network slices and applications.

To Qualcomm's questions:

Thanks for the questions.

DRB related L2 metrics can be queue size, data transmitted in a time interval, etc. They can be measured based on the data volume or buffer size of MAC/RLC/PDCP entities associated with a DRB. It is similar to data volume calculation defined in RLC/PDCP specifications for buffer status report defined in MAC specification. The intention is to allow such L2 metrics to be visible to SDAP entity, similarly to RLC/PDCP data volume being visible to MAC entity, when QoS flow to DRB mapping is performed.

To Apple's question:

Thanks for the question.

As scheduling/grant and LCP operation in MAC, SDAP operation only directly impacts transmission at UE; they are not intended for QoS control at other network nodes. End-to-end QoS measurement and control require all network nodes along the communication path to perform their respective role. SDAP entity provides the means to perform QoS control at a UE, based on the QoS requirements derived for the segment from the UE to a network node (either a gNB, an IAB node, or a relay node). QoS measurement and enforcement at Relay node is a different issue, which has been discussed separately in Rely architecture, e.g., if and how to configure a relay node to be aware of and enforce QoS requirements for traffic of individual UEs. Relay deployment is also less common for latency critical applications.

3.4 Round 2 Questions on RWS-210040

Feedback Form 6: Round 2 Questions on RWS-210040

1 – VODAFONE Group Plc

Thanks for the interesting paper. However, why can't the simple "LTE-like" solution of mapping one QoS flow to exactly one DRB and Logical Channel Group be used (especially as NR has more LCGs than LTE has)?

3.5 Round 2 Answers on RWS-210040

To Vodafone's question:

Thanks for the question.

The premise of 5G QoS framework is to support much more diverse applications on a UE, possibly from multiple PDU sessions and network slices, without a need of dedicating a DRB to each application. There is a maximum number of DRBs a UE is required to support, given the processing and memory resource required when a DRB is configured. Hence, the one-to-one mapping relation between EPS bearer and DRB of LTE is not adopted in 5G, when 5G Core took up QoS flow approach. It is common understanding that multiple QoS flows may be mapped to one DRB to practically support a rich set of applications in 5G without requiring UE to be overly dimensioned.

Even with one DRB dedicated to one QoS flow, current UL grant and LCP procedures are still not suited to deal with traffic burst, e.g., to enforce MDBV (Maximum Data Burst Volume) and Slice MBR. As discussed in RAN2, LCP operation, once configured by RRC signaling, wouldn't adjust to the data burst of a QoS flow, e.g., stopping transmitting this QoS flow's data if its MDBV or Slice MBR are being exceeded at a time instant.

Thanks for the question.