3GPP TSG-RAN3 #112-e R3-212690

17th – 27th May 2021

Online

Agenda Item: 18.4

Source: ZTE (moderator)

Title: SoD Standards Impact on Existing Nodes, Functions, and Interfaces

Document for: Approval

# Introduction

**CB: # 47\_DataColl\_StdImpact**

**- Chair: suggest to structure discussion around 6 areas depending on the use cases agreed in CB 46, splitting work among companies for the resulting TPs, e.g.:**

**1) Common (merging any agreeable parts from e.g. 1684, 1858, 1970, 2034) (ZTE,CU?)**

**2) ML models (merging any agreeable parts from e.g. 2371, 2374, 1755) (QC?)**

**3) Load balancing (merging if agreeable from 2505, 2316) (CMCC?/E///?)**

**4) Mobility (merging any agreeable parts from e.g. 2465, 2546, 2180) (SS?)**

**5) ES/EE (merging if agreeable from 2315) (E///)**

**6) Others (merging if agreeable from 2525) (HW)**

**- Chair: NB the outcome of this discussion is a direct consequence of CB 46, i.e. we should discuss and agree use cases first, and then consider the impacts for the agreed ones.**

(ZTE - moderator)

# For the Chairman’s Notes

Propose the following:

**AI-enabled RAN architecture:**

**Proposal 1: Agree to capture AI-enabled RAN architecture into the TR37.817. Take the TP section of R3-212034 as baseline and revised in R3-21xxxx.**

**Interface enhancement for AI:**

**Proposal 2: WA: The common interface enhancement for AI should support AI Function Management (Start/Stop AI function) and AI Model Management (Model Distributing/Updating).**

**Proposal 3: Security aspect should be considered and coordinate with other working groups later if needed.**

**ML training/ML inference deployment:**

**Proposal 4: Model training and Model inference can be deployed into different places, and where model training/model inference is located should be studied case by case.**

**Standard impacts on specific use case:**

**Load balancing/Load prediction:**

**Proposal 5: AI-based load prediction should be regarded as a separate use case to discuss the corresponding standard impact.**

**To be continued:**

**The detailed standard impacts for load balancing/load prediction should continue to discuss at next meeting.**

**Majority companies prefer predicted load exchanged between peer NG-RAN nodes should be supported.**

**For load prediction, historical radio resource status can be the input and predicted resource status of serving cell and neighbor cell can be the output.**

**Mobility optimization:**

**Proposal 6: AI-based trajectory/location prediction should be regarded as a separate use case to discuss the corresponding standard impact.**

**To be continued:**

**The detailed standard impacts for mobility optimization should continue to discuss at next meeting.**

**For trajectory/location prediction, UE mobility history information can be the input, and UE trajectory prediction or the predicted target cell can be the output.**

**Energy saving:**

**Proposal 7: Load prediction/trajectory prediction can be used for energy saving decision, and should be regarded as a separate use case to discuss the corresponding standard impact.**

**To be continued:**

**The detailed standard impacts for energy saving should continue to discuss at next meeting.**

# Discussion

This CB focuses on discussing the standards impact on existing nodes, functions, and interfaces.

## AI-enabled RAN architecture

The AI-enabled RAN architectures (involving split architecture and non-split architecture) are proposed in [4] as:

 

Companies are invited to provide their comments on the AI-enabled RAN architecture

|  |  |
| --- | --- |
| **Company** | **Comments** |
| ZTE | Taking the general objective as the studies should be focused on the current NG-RAN architecture and interfaces to enable AI support for 5G deployments into account, the possible AI-based RAN architecture (involving split architecture and non-split architecture) are shown above.  Data Collection and Action are usually should be located in NG-RAN node for those use cases identified for RAN optimization. While Model Training and Model Inference can be both located in a single place, e.g. OAM system or NG-RAN node, or Model Training is located in the OAM system, and Model Inference is located in the NG-RAN node. How to define the functional components of the AI Entity and the deployment of AI Entity should be discussed case by case.  Capture the AI-based RAN architecture into the TR is helpful for further discussion. |
| Nokia | We support the AI-enabled RAN architecture illustrated in the figures above. In principle we can have that Training and Inference are both located in a common entity (e.g., in cases of Reinforcement Learning) or in different entities (e.g., in case of Supervised Learning). In our view the following options are possible: a) Model Training and Model Inference in OAM, b) Model Training in OAM and Model Inference in the RAN, and c) Model Training and Model Inference in the RAN. Those 3 alternatives are captured by the above figures which show in a general way (applicable for all use cases) the possible placements of the AI/ML entity. |
| Deutsche Telekom | The architecture diagram is generally fine with us, but we have some comments for improvements:   * To have a complete view, an AI entity should be also added to the 5GC. * With current representation of AI entities, there is the impression that they are separate to RAN nodes or OAM. From our perspective they should be part of a RAN node or OAM domain, i.e. the corresponding blocks should go around the AI entities.   We agree with the statements of ZTE and Nokia w.r.t. deployment alternatives for Model Training and Model Interference, but there is need to differentiate between offline and online training. We see generally the initial offline training as part of OAM domain. |
| Samsung | It has been agreed that the study is based on the current architecture and interface. We believe the NG-RAN nodes have the function of AI. But detailed deployment and standard impact should be studied case by case. |
| Intel | We support to capture the above figure of AI-enabled RAN architecture in the TR. Following comments suggest being considered for improvement and clarification:  1. Agree with DT, for the completeness, we also suggest adding an AI entity in 5GC.  2. The location of model training and model inference for each AI entity can be discussed case by case. All AI entities in different network nodes (i.e. OAM, NG-RAN, gNB-CU, gNB-DU) may have the capability of ML training and ML inference.  3. We suggest using Figure 6.1-1 in TS38.401 as the baseline, where OAM and AI entities can be added in the figure. It is because Figure 6.1-1 in TS38.401 includes both gNB and CU-DU split architecture in the same figure. |
| vivo | We think the AI-enabled RAN architecture is essential for use case discussion. |
| NEC | Having RAN AI architecture or deployment scenarios in TR could be beneficial.  Regarding the architecture above, one point. Some papers mention that ML model training could be in CN/OAM. Why figures above do not show AI entity in 5GC? |
| Futurewei | The AI-enabled RAN architecture is ok with us.   * We agree with Nokia that three potential learning scenarios can be supported by the architecture, however, it is not necessary training and inference have to reside on the same entity in RL case. * We agree with DT that AI entity/module should be part of RAN node. |
| CATT | Agree with this figure. |
| Huawei | Not sure if we should do this way. As we could see that we never put RRM or SON everywhere in the RAN node, to introduce an entity may also mislead to the conception of a standalone entity which actually is not in line with consensus.  Of course, some function descriptions are necessary.  Then for the location of each function, as discussed in our paper, in general, we think action is anyway conducted in RAN area while for collection, training and even inference, could be located in OAM which may also be case dependent. |
| CMCC | We agree with to capture the architecture in the TR. We may consider whether AI entity is appropriate, since it is a function of RAN node. |
| AT&T | Agree to capture both architecture options, but we are not sure if “AI entity” is the best wording as it could represent AI functionality split across multiple nodes managed by a common “entity” depending on the actual deployment scenario. |
| LGE | Better to capture from function point of view. Share the concern of DT, Samsung and Huawei |
| Ericsson | Agree with Huawei about the fact that AI is a function that can be supported by specific logical nodes. The diagram up for agreement is a logical nodes and interface diagram, where we have never expressed functionalities. For that wed not think that “AI” should be marked in an architecture diagram of that kind.  Instead, we would suggest to express where the AI functionality is supported both in the functional framework and in the use case description.  We are fine with the concept of having inference in the RAN and Training in the RAN, or Inference in the RAN and training in the OAM. For the concept of inference and training in the OAM this is also a valid scenario, but it goes beyond the scope of work in RAN3. This is because the scenario would totally rely on operations in the OAM, which should be rather discussed in SA5.  We need to keep in mind that our standardization impacts should be determined on a use case basis. The WID in RP-201620 mentions:  *Study standardization impacts for the identified use cases* |
| Lenovo, Motorola Mobility | The figure looks in general fine. On the other hand, if we assume AI entity can be located in each RAN node, would it be enough to just mention it in the spec? |
| Qualcomm | Agree in principle.  The AI entity in the figure should be primarily for inference and training. The data collection and storage should be separate entity. |

**Moderator’ s summary:**

Majority of companies agree to capture the AI-enabled RAN architecture into the TR. Among them, some clarification and improvements need to be considered (e.g. AI entity in 5GC, deployment of training and inference and wording of AI entity), while some companies think that AI entity should not be introduced into the diagrams which causes misunderstanding.

In order to reach a consensus, the AI-enabled RAN architectures above are modified for agreement as follows:



Figure 1. AI-enabled RAN architecture

**Proposal 1: Agree to capture AI-enabled RAN architecture into the TR37.817. Take the TP section of R3-212034 as baseline and revised in R3-21xxxx.**

## Interface Enhancement for AI

[4][6] proposed to enhance the existing interfaces in the RAN to support the AI function.

The proposals:

**The interface enhancement for AI is to support:**

* **AI Function Management: Start/Stop AI function.**
* **AI Model Management: Model distribution/updating**
* **AI Measurement Management: AI measurement report exchange**
* **Mechanisms to enable and disable ML operations on a per-need basis**
* **Mechanisms to start and stop ML predictions**

Companies areinvited to provide their comments on these proposals above.

|  |  |
| --- | --- |
| **Company** | **Comments** |
| ZTE | AI related interface functions need to be defined and discussed further:  - AI Function Management: This function enables the AI function(s) start/stop between the NG-RAN nodes, if the AI function(s) supported by both nodes.  - AI Model Management: This function enables one NG-RAN node to retrieve the ML model from the peer NG-RAN node. With this function, the updated ML model can be synchronized between NG-RAN nodes.  - AI Measurement Management: This function allows AI measurement report between NG-RAN nodes.  And for the last two bullets, these to mechanisms can be included in the AI Function Management. |
| Nokia | AI/ML functionality should be activated on a per need basis. For this reason, we see the need to have AI Function Management introduced. However, it should also provide means to neighbouring nodes to determine if the AI Function is supported by a neighbour. In our view there are two aspects of AI Functionality at a node. One is related to a capability to run ML at a given point in time since it may conflict with other internal operations of the node. But furthermore, once this capability is present, it may further refer to activating/deactivating ML models for specific use cases (e nergy saving, load balancing, etc.) since ML may not be active for all use cases at a given time if baseline operation (not using ML) has good enough performance.  It is not clear to us from the description what are the details of AI Model Management. However, we could think that the network should be able to forward an ML Model to the entity responsible for executing it. We cannot think why Model synchronization is necessary at this stage.  AI Measurement Management is also important even though in our view it is not only necessary for AI Report exchange but also for handling predictions and other ML-based information.  We see the need that different gNBs can request from each other predictions in a coordinated way. This would mean that an NG-RAN node should be able to request predictions from a peer NG-RAN node and the latter should report this information when possible. |
| Deutsche Telekom | AI management functionalities should be based on a SBA-like approach. AI functions in a RAN node may request/subscribe services from other functions in other nodes during a certain time span (like data collection & preparation based on measurements, data analytics/predictions, model exchange/update etc.). |
| Samsung | Specific interface impact needs to be studied case by case, so it seems unnecessary to define the general interface function.  And it is unclear about the definition of AI measurement management. Still a little bit confused about what is AI measurement report. |
| Intel | We agree with DT that AI management should be service based.  Beyond the above listed interface enhancement, we also think data management procedure should also be considered, such as data request, data response, etc. |
| vivo | We generally agree with the intention. While the enhancement for specific interface should be discussed case by case. |
| NEC | We believe that priority should be given to data collection for ML model training/validation/inference and to outputs from ML model to NG-RAN nodes.  But topics mentioned above could also be studied. |
| Futurewei | It is reasonable to support AI Function Management and AI Model Management.  It is not clear what “AI Measurement Management: AI Measurement Report exchange” represents. This item needs to be clarified first.  The last two items (“Mechanisms to enable and disable ML operations on a per-need basis” and “Mechanisms to start and stop ML predictions”) should be part of the AI Function Management. |
| CATT | We really appreciate DT’s view that an SBA-like approach is much better, but it seems contradicting the principle of “do not define new interface” in the SI document…  And for the functions needed to be supported, we agree with all the 5 bullets listed by the moderator.  For the definition raised by ZTE, we think the first bullet and the second bullet needs further discussion. E.g. exchanging models between peer nodes doesn’t seem very useful for any use case. |
| Huawei | In our understanding, any discussions about spec impacts should be coupled with use case, but not to have some procedural actions like start/stop indication.  Then I am not sure such function management/model management/measurement management is really needed or not, we have interface management procedure, mobility management procedure or session management procedure, and AI/ML actually will serve as a tool to achieve e.g. mobility function. |
| CMCC | First, apart from impacts on RAN internal interfaces, e.g., Xn/F1/E1， impacts to the interface between RAN node and OAM should also be considered, RAN3 can study the requirement on this interfaces.  For the interface impacts, first, we admit the trend from DT that SBA approach can be considered. However, we noticed that companies have put forward SBA based NG interface in Rel-18 in SA2, but companies are still hesitate to consider it. So for the RAN interface, at current stage considering SBA approach may moving too far.  For the function of interfaces, in general, we think at least **Model distribution/updating an ML results, e.g., prediction results based on ML inference, can be supported** |
| LGE | We need to check the use cases one by one and decide the impacts to the interfaces. Case by case approach is preferred. |
| Ericsson | We ned to remind that the work on AI is tructureed as follows, as per SID in RP201620:   * 1. *Study standardization impacts for the identified use cases including: the data that may be needed by an AI function as input and data that may be produced by an AI function as output, which is interpretable for multi-vendor support.*   2. *Study* *standardization impacts on the node or function in current NG-RAN architecture to receive/provide the input/output data.*   3. *Study standardization impacts on the network interface(s) to convey the input/output data among network nodes or AI functions.*   *One general objective for the work is that* *the studies should be focused on the current NG-RAN architecture and interfaces to enable AI support for 5G deployments.*  Hence, the approach to identify standardization impacts needs to be on a case by case basis, as for example showed in R3-212315 and R3-212316.  Regarding the proposed interface enhancements:   * **AI Function Management: Start/Stop AI function**   We do not see the need of this function, just like we do not have a SON start/stop function. Functions are deployed in nodes that support them by means of configuration. Their activity to external nodes is visible via procedures over the interfaces.   * **AI Model Management: Model distribution/updating**   As discussed in CB 45, defining a model distribution/updating procedure over common interfaces implies to define the model itself as an AP IE. This contradicts the agreement taken in RAN3 that the model is implementation specific. Therefore, a function for model transfer should be out of RAN3 scope, unless RAN3 decides to transfer a model as a transparent container, in which case the procedure becomes intra vendor.   * **AI Measurement Management: AI measurement report exchange**   The need for measurement reports should be evaluated on a case by case basis. Measurement reports are exchanged today over common interfaces in support of several functions like mobility, interference management, load control. Not for this we lable every measurement exchanged over the interfaces with a function name. Hence, once the measurements needed for the AI use cases are identified, we can simply specify their transfer over the interfaces without the need of an AI label to it.   * **Mechanisms to enable and disable ML operations on a per-need basis**   Again this should be studied on a case by case basis. Rather than “enabling or disabling” ML operations, the work in RAN3 has focused on enabling/disabling exchange of information between nodes, in support of certain functions. Enabling of a function in a node is a configuration related procedure outside of RAN3 scope   * **Mechanisms to start and stop ML predictions**   We interpret this as a mechanism to allow different nodes to receive predictions. Such predictions may be ML generated…this is not really important. We agree to this principle of “subscribing to ML outputs”. Namely, a nod can request to receive certain predictions for a different node and with certain reporting characteristics. The same node can stop the reporting if needed.  Note1: just like a mechanism of “subscribing to ML outputs” is needed, a mechanism of “subscribing to ML inputs” would also be needed. Namely, a node needs to be able to request reporting of data that the ML function hosted by the node will use as inputs.  Note2: the need for subscription based inputs/outputs needs to be studied on a case by case basis |
| Lenovo, Motorola Mobility | As start point, RAN3 is suggested to focus on the inference data/training data/inference output/measurement result exchange among relevant nodes. Also, the training data provision is relevant to the model update, e.g. when the model update is triggered, the relevant training data shall be provided to the ML training.  If time allows, RAN3 can further the AI model distribution (we consider this is highly dependent on the exact AI algorithm, e.g. if it is a federated learning). |
| Qualcomm | Agree in general.  What is “AI measurement”?  ZTE: AI measurement management is to allow measurement report between different nodes (e.g. required data for training/inference request/response). |

**Moderator’s summary:**

8 companies deems that these common interface enhancement for AI (e.g. AI function start/stop, model distributing/updating) should be supported, while 6 companies prefer that the interface enhancement for AI should be discussed case by case later.

**Proposal 2: WA: The interface enhancement for AI should support AI Function Management (Start/Stop AI function) and AI Model Management (Model Distributing/Updating).**

In [12], it is proposed to consider security issue for data accessibility, and training, data collection and inference could be operated in different place. Authorization of data accessibility to access data repository services shall be supported in RAN.

Companies areinvited to provide their comments on security issue for data accessibility.

|  |  |
| --- | --- |
| **Company** | **Comments** |
| ZTE | The security requirement on AI related data over interfaces should be considered, and the robustness and reliability of AI related data transmission should also be guaranteed. This can be discussed further with contribution driven. |
| Nokia | Security is of course an important aspect that we should always keep in mind. However, it is not clear to us why this aspect will be different for AI/ML as opposed to normal baseline data collection. We could check with SA3 if needed. |
| Deutsche Telekom | Security aspects should be considered in the study. A cross-check with SA3 can be triggered after availability of more details on use cases and architectural impact. |
| Samsung | Security should be considered and can be discussed further with contribution driven. |
| Intel | Agree with Nokia and DT, we should cross-check with SA3 once we have more mature understanding what data is collected and requested by AI functions. |
| vivo | Agree with Nokia and DT that SA3 should be involved after the discussion about framework and use case are converged. |
| NEC | This topic is important for AI/ML because of the amount of data collected and potential possibilities to extract various information from the collected data.  However, not sure there is enough time for this topic in this study item. Maybe for later stage in coordination with SA3. |
| Futurewei | Security aspects should be investigated for data accessibility, and any information exchanges over the interfaces for AI/ML related functions. This security aspect should be consistent with existing security handling of data collection and information exchanges. |
| CATT | Agree with Nokia, DT and Intel. |
| Huawei | Security of cause is important, and as suggested in the discussion paper, this should be discussed in SA5 since all the data are stored in OAM, thus we think an LS is needed to indicate RAN3 consensus. |
| CMCC | Agree with Nokia, DT and Intel. |
| LGE | Important, can be studied and coordinate with SA3 on the identified specific issues |
| Ericsson | There is a need for more discussion on why security is different for AI than for e.g. SON or MDT. However, RAN3 does not seem to be the place to discuss such topics, so to prioritise our work in a correct manner we should probably not spend time on this topic and rather trigger the topic in other groups like SA3 or SA5. |
| Lenovo, Motorola Mobility | The question is whether we need security enhancements to support AI functionalities. It can be dependent on exact case, e.g. if ML model is exchanged. We can discuss case by case. SA3 and SA5 should be involved in for further study. |
| Qualcomm | Agree. Security is very important for data access. We can cooperate with SA3 on this. |

**Moderator’ s summary:**

It seems that all companies deem that security for data accessibility should be considered to study, and most companies think SA3 or SA5 should be involved in later for further study.

**Proposal 3: Security aspect should be considered and coordinate with other working groups later if needed.**

## Model training/Model inference deployment scenarios

This section is intended to discuss where Model training/Model inference should be located. Some companies think Model training and Model inference can be deployed into different places according to the use case, other companies think ML training should be located in the OAM rather inside RAN. Hence, companies are invited to provide their understanding and comments on the below proposals:

1. Model training and Model inference can be deployed into different places according to the use case.
2. ML training should always be located in the OAM,

|  |  |  |
| --- | --- | --- |
| **Company** | **Which prefer?**  **(P1 or P2)** | **Comments** |
| ZTE | P1 | Model Training and Model Inference can be both located in a single place, e.g. OAM system or NG-RAN node, or Model Training is located in the OAM system, and Model Inference is located in the NG-RAN node. How to define the deployment of AI Entity should be discussed case by case.  For example, for AI based Energy saving, the Model Training can be located in the OAM system. For Trajectory Prediction, both the the Model Training and Model Inference can be located in the NG-RAN node.  For UE trajectory, AI-based trajectory prediction may be to predict the UE trajectory on the seconds granularity, so if Model training and Model inference are both deployed at OAM, required measurements report from UE or NG-RAN node have to be transferred to OAM for training and inference. However, data transmission over backhaul interface cannot meet the requirement on real time. It’s not a suitable solution to the AI-based trajectory prediction for making mobility decision which needs to be adjusted in real time.  For AI-based load prediction, predicting the traffic load is on the days or hours or minutes granularity. Different from the solution to real-time use case, Model training and Model inference can be both deployed in the OAM for a long-term time scale, and inference stage can also be located in the NG-RAN node for the short-term predicted load. |
| Nokia | P1 | It should be possible for Training and Inference to be located at different places according to the use case. Offline training could be located in OAM but online training should be in the RAN. |
| Deutsche Telekom | P1 | Same view as Nokia, but we see initial offline training only in OAM (“shall” instead of “could”). |
| Samsung | P1 | The deployment should be studied case by case with the consideration of node computing capability, data availability, output utilization, etc. |
| Intel | P1 | As we commented in Q1, the location of model training and model inference for each AI entity is possible to be allocated to different network nodes (OAM, NG-RAN, CU, DU). This can be discussed case by case. We do not need to limit the deployment scenario. |
| vivo | P1 | Whether OAM is proper place is based on the existing data in OAM and capability of OAM, i.e., if the training data already exist in OAM and the OAM is able to perform the training, then OAM is proper. Otherwise, some other solutions are not precluded. |
| NEC | P1 | When use cases and solutions are more or less stabilized, it would be possible to summarize deployment scenarios.  Also, maybe better to concentrate on deployment options of ML model independent parts. For example, as was mentioned in some papers interfaces between Model Training and Model Inference are ML model dependent, which basically means that they are implementation dependent, at least within the current scope of this study item. On the other hand, interfaces between Data Collection and Model Training and between Model Inference and Actor could be standardized and thus it is important to study various option of mapping them into RAN interafces. |
| Futurewei | P1 | As discussed in section 3.1 AI-enabled RAN architecture, there may be three training/inference scenarios: a) Both Model Training and Model Inference in OAM,  b) Model Training in OAM and Model Inference in the RAN,  c) Both Model Training and Model Inference in the RAN  P1 is able to support all three scenarios. |
| CATT | P1 | P2, i.e., making everything at OAM, will result in huge and persistent signalling load between NG-RAN and OAM. |
| Huawei | both | We think P1 and P2 are not contradictory with each other, no? i.e. training should be always located in OAM, while inference could be also in OAM (for example, energy saving which doesn’t require real time action), or in RAN which could be close to the actor where real time action is needed. |
| CMCC | For sure P1 |  |
| AT&T | P1 |  |
| LGE | P1 | Should be open to cover all the use cases |
| Ericsson | P1 with caveat | P1 includes P2. We agree that model training could be placed in the RAN or in the OAM. One implication of agreeing to the possibility of separation between the training and the inference functions is that there will be the need to transfer the AI/ML model. However, as discussed in CB45, the model is a black box that cannot be openly encoded.  Hence the caveat to accepting P1 is that model transfer in case of training and inference functions in different entities is proprietary. |
| Lenovo, Motorola Mobility | P1 with comment | Where the ML training and ML inference are located shall be analyzed case by case.  On the other hand, we believe it is beneficial to have a baseline assumption (e.g. where ML training and ML inference are located) when dive into case by case analysis. In other word, “everything is possible” mindset is not helpful to make progress. |
| Qualcomm | P1 | Online training can be performed in any network entity depending on the data availability. |

**Moderator’s summary:**

Basically, most companies prefer P1 that Model training and Model inference can be deployed into different places according to the use case.

**Proposal 4: Model training and Model inference can be deployed into different places, and where model training/model inference is located should be studied case by case.**

## Standard impacts on specific use case

According to the chairman notes “**NB the outcome of this discussion is a direct consequence of CB 46, i.e. we should discuss and agree use cases first, and then consider the impacts for the agreed ones.**” , in this section, we should focus on standard impacts the agreed use case depending on the CB#46.

### Standard impacts for AI-based load balancing (if agreed)

Companies are invited to provide their views on the solution and standard impacts for AI-based load balancing.

|  |  |
| --- | --- |
| **Company** | **Comments** |
| ZTE | In our understanding, load prediction can be a tool-box use case for the load balancing. Predicted load can be the augmented information for network to make load balancing decision.  The corresponding standard impacts of load prediction use case may include:   * AI Measurement management procedures between NG-RAN nodes in order to report load prediction required data and the corresponding assistance information, e.g, Historical traffic load of cells; * AI function management procedures in order to align the corresponding AI functions between NG-RAN nodes; * AI Model management procedures in order to distribute/update the ML model between NG-RAN nodes; * The predicted load information is exchanged with the neighbour NG-RAN nodes via Resource Status Report procedure which can be used for subsequent optimization (e.g. energy saving, load balancing, mobility management, etc). |
| Nokia | We agree with the idea of having load predictions available in the RAN but maybe we can reword from “tool-box use case” to support of a basic load prediction functionality for AI/ML. We do support to introduce mechanisms that will enable peer RAN nodes to exchange load prediction information. Using Resource Status Report procedure is a valid option. |
| Deutsche Telekom | We should first focus on detailed use case descriptions (see CB # 46) before going into details of standard impact. |
| Samsung | There are two types of AI-based load balancing:   * Type1 ML-assisted load balancing: Load balancing strategy is generated by conventional method based on the predicted resource status, where the predicted resource status is generated by ML model. AI functionality is to generate predicted resource status. * Type2 ML-generated load balancing: Load balancing strategy is generated by ML model based on the current/historical resource status. AI functionality is to generate predicted load balancing strategy.   So NG-RAN node is required to support ML model inference, exchange predicted resource status or predicted load balancing strategy. |
| Intel | Besides the load prediction information, we further propose to include the predicted resource status between different NG-RAN nodes. With the predicted resource status from the target NG-RAN, load balancing handover decision at the source NG-RAN node can thus prevent handover unsuitable workload which may lead to overload at the target NG-RAN node.  Either to use Resource Status Report or having a new message transmitting all predicted information can be FFS. |
| vivo | Agree with DT that the standard impacts shall rely on the conclusion of CB #46 |
| NEC | We believe that Network Load Balancing and Load Prediction are different use cases and should be considered and described as two different use cases.  The goal of Network Load Balancing is to achieve more balanced distribution of load among cells.  The goal of Load Prediction is to generate time series of predicted load values for a set of cells.  Output of Load Prediction could be used as one of the inputs to Network Load Balancing. |
| Futurewei | Agree with DT and vivo. Standards impact for AI-based use cases should be discussed after the description and details for each use case are agreed (under CB #46). |
| CATT | We support the idea of load prediction, i.e. what ZTE means, and as well as Type1 in Samsung’s comment.  For Type2 in Samsung’s comment, we think we should first identify whether it is really beneficial. |
| Huawei | For solution, as indicated in our discussion paper in R3-212524, the cell load information (may include load prediction info) could be exchanged between nodes, and also UE trajectory prediction info could also be used. |
| CMCC | As shown in our paper, at least introducing mechanisms that will enable peer RAN nodes to exchange load prediction information should be studied (i.e., for type 1 in Samsung terminology) and the flow chart should be captured in the TR  For the type 2 solution, we can also study it and find solutions |
| LGE | After the use cases are agreed, then we can check the parameters necessary. |
| Ericsson | Here is a description of the load balancing use case as included in R3-212326:   * A RAN node can request and obtain UE traffic predictions, based on real end-user behaviour. This provides RAN with insights on the traffic that UEs would generate in the future. This mechanism implies configuring a UE to collect and report such prediction over RRC protocol, RAN2 needs to be consulted for that. * A RAN node can also predict its own load. This can be achieved by considering the own load, load information received from neighbor RAN nodes, and traffic predictions from UEs. Finally, RAN predictions can be signaled between RAN nodes. In addition to the signalling impact foreseen on RRC protocol, this second mechanism has an impact on RAN3 protocols, such as XnAP, X2AP, F1AP, E1AP to be discussed during the normative phase.   A high level signaling flow for the case described is shown in the figure below. |
| Lenovo, Motorola Mobility | As start point, RAN3 is suggested to focus on the inference data/training data/inference output/measurement result exchange among relevant nodes. Also, the training data provision is relevant to the model update, e.g. when the model update is triggered, the relevant training data shall be provided to the ML training.  If time allows, RAN3 can further the AI model distribution (we consider this is highly dependent on the exact AI algorithm, e.g. if it is a federated learning). |
| Qualcomm | The standard impacts details should be decided based on CB#46 conclusion.  The load prediction should be one aspect of standard impact at least. It could include standardized input and output parameters for different kind of predictions. |

**Moderator’s summary:**

5 companies think the standard impacts for AI-based use cases should be discussed after the description and details for each use, while a number of companies deem that load prediction can be used for AI-based load balancing, and load prediction should be regarded as a separate use case to discuss the corresponding standard impact. Some companies also think predicted load exchanged between peer NG-RAN nodes should be supported.

**Proposal 5: AI-based load prediction should be regarded as a separate use case to discuss the corresponding standard impact.**

**To be continued:**

**The detailed standard impacts for load balancing/load prediction should continue to discuss at next meeting.**

**Majority companies prefer predicted load exchanged between peer NG-RAN nodes should be supported.**

Companies are invited to provide their views on the input/output for AI-based load balancing.

|  |  |
| --- | --- |
| **Company** | **Comments** |
| ZTE | For load prediction:  The following data is required as input data for load prediction.   * Historical traffic load of cells; * Historical PDCP Data Volume of cells; * Historical PRB utilization rate of cells; * Historical RRC connection number of cells.   The output of load prediction contains the following information.   * Trends of the traffic load * Predicted load of serving cell and neighbour cells |
| Nokia | Input to the ML Load Balancing function running at a RAN node may be predicted load information requested from its neighbouring peer nodes. Evaluating the need for exact measurements as an input to Load Balancing (e.g., historical RRC connection number of a cell, historical PRB utilization rate) is more complex and difficult since ML models are not in the scope of the SI and different implementations may have different required inputs. A definition of “historical” information is also needed.  Output of ML Load Balancing function can be the  -Expected load of a served cell  -Expected load to be transferred to a neighbouring cell |
| Deutsche Telekom | We should first focus on detailed use case descriptions (see CB # 46) before going into details of standard impact. |
| Samsung | The input for both type 1 and type 2 can include   * node resource status * neighbor resource status   The output can be:   * Type 1: predicted resource status, inc.:   + predicted resource status   + predicted resource increment/reduction   + predicted status corresponding time * Type2: predicted load balancing strategy, inc.:   + action type, inc. transferring out, transferring in, etc.   + source node   + target node   + load transferring amount   + action time |
| Intel | Input data may also require below information:   * Historical radio resource status * Historical TNL capacity   Output may include:   * Predicted load of a served cell * Predicted resource status of a served cell * Accuracy of the prediction result * Validity time of the prediction result |
| vivo | Agree with DT that the standard impacts shall rely on the conclusion of CB #46 |
| NEC | Output of Network Load Balancing is decisions/recommendations for a set of cells to re-distribute load, for example, by performing load-balancing HOs.  Output of Load Prediction is time series of load values for a set of cells.  Output of Load Prediction could be used as one of the inputs to Network Load Balancing. |
| Futurewei | Agree with DT and vivo. Standards impact for AI-based use cases should be discussed after the description and details for each use case are agreed (under CB #46). |
| CATT | For load prediction, the output should be the predicted load of cells. And if the inference module is at an entity within RAN, not only the serving cell but also the neighbor cells can be subject of prediction, as pointed out by many companies.  The accuracy of prediction may also be provided, although not very essential for load prediction we think.  For inputs of prediction, we general agree with the parameters listed above. Some other inputs may also be included after further study, especially after simulation. |
| Huawei | * Input:   + UE mobility history info, if needed, to the NG-RAN node 1 at the source cell.   + UE performance measurement at the target cell, e.g. E2E delay, if needed.   + Neighbour cell load prediction from the NG-RAN node 2 to the NG-RAN node 1.   + Reward information from the NG-RAN node 2 to the NG-RAN node 1 after the load balancing actions. * Output   + Load balancing decision, e.g. HO command |
| CMCC | For the load prediction case, similar view as intel  Input data :   * Historical radio resource status and other load metric has been defined in SON/MDT WI   Output may include:   * Predicted load of a served cell * Predicted resource status of a served cell * Accuracy of the prediction result * Validity time of the prediction result |
| Ericsson | Possible inputs to the load balancing AI model could be:   * Assistance information from the UE revealing possible traffic patterns * Load Metrics currently specified in Resource Status procedures, for both serving RAN and neighbour RAN nodes   Possible outputs generated by a load balancing model   * Prediction of load metrics (e.g. metrics defined in the Resource Status procedures) for the serving RAN, which can be signalled to neighbour nodes. * Predictions of load metrics (e.g. metrics defined in the Resource Status procedures) for neighboring cells |
| Qualcomm | Agree with DT, Vivo and Futurewei. The details can be decided later on.  The load prediction can be performed for different nodes, different resource and traffic.  For each prediction, history load and current load are used as input for sure. The output should include the prediction result and confidence level. |

**Moderator’s summary:**

4 companies think the standard impacts for AI-based use cases should be discussed after the description and details for each use, while a number of companies mentioned that the load prediction is benefit for load balancing decision. For load prediction, historical radio resource status can be the input and predicted resource status of serving cell and neighbor cell can be the output.

**To be continued:**

**For load prediction, historical radio resource status can be the input and predicted resource status of serving cell and neighbor cell can be the output.**

### Standard impacts for AI-based Mobility optimization (if agreed)

Companies are invited to provide their views on the solution and standard impacts for AI-based mobility optimization.

|  |  |
| --- | --- |
| **Company** | **Comments** |
| ZTE | In our understanding, trajectory prediction can be a tool-box use case for the mobility optimization. Predicted UE trajectory can be the augmented information for network to make mobility optimization decision.  Hence, we discuss standard impacts for UE trajectory prediction here:  The corresponding standard impacts of UE trajectory prediction may include:   * MDT signalling enhancement in order to report UE history location information and the corresponding assistance information, e..g, UE moving direction, UE velocity, Radio measurements; * AI function management procedures in order to align the corresponding AI functions between NG-RAN nodes; * AI Model management procedures in order to distribute/update the ML model between NG-RAN nodes; * The UE trajectory prediction information may be included in the handover request message towards the target NG-RAN node, which can be used for subsequent optimization. |
| Nokia | We agree with the idea of having trajectory predictions available in the RAN but maybe we can reword from “tool-box use case” to support of a basic trajectory prediction functionality for AI/ML. UE trajectory prediction information at a node may help it to better prepare Handovers (or Target Cells for Conditional Handovers).  Even though we think that MDT could be a useful procedure to provide the network with UE measurements, it will be difficult to study MDT enhancements without RAN2 involvement in this SI, especially when it comes to introducing new UE measurements. |
| Deutsche Telekom | We should first focus on detailed use case descriptions (see CB # 46) before going into details of standard impact. |
| Samsung | ML model can help to generate the adaptive handover strategy directly to keep or enhance the QoS performance during HO by improving the HO efficiency and reducing the HO failure.  So NG-RAN node is required to support ML model inference.  And we also agree with the idea of trajectory prediction to enhance mobility management, and NR-RAN to exchange predicted trajectory for subsequent optimization. |
| Intel | We think UE trajectory prediction use case is different with mobility enhancement. By looking at the use case description in CB #46, mobility enhancement is trying to collect UE performance for successful handovers and trying to optimize future handover decision, such as threshold setting, cell selection, DC configuration, etc.  We agree with the intention that UE predicted location can help to improve performance of handover. However, this should be considered separately. |
| vivo | Agree with DT that the standard impacts shall rely on the conclusion of CB #46 |
| NEC | We believe that Mobility Optimization and UE Position Prediction are different use cases and should be considered and described as two different use cases.  The goal of Mobility Optimization is to increase number of successful HOs and decrease number of unsuccessful HOs (e.g., HOs resulted in RLF).  The goal of UE Position Prediction is to generate time series of predicted coordinates of a UE or of a group of UE.  Output of UE Position Prediction could be used as one of the inputs to Mobility Optimization. |
| Futurewei | Agree with DT and vivo. Standards impact for AI-based use cases should be discussed after the description and details for each use case are agreed (under CB #46). |
| CATT | We think both AI-based trajectory prediction and AI-based HO (and potentially CA/DC configuration) decision are useful. Both functions seem complex. |
| Huawei | The whole procedure is quite similar, here the UE trajectory prediction info, radio quality measurement and cell load info might be used as input for ML from which there will be a mobility decision or policy for mobility decision.  After handover the target node may send the reward information to the source node, the source node could update the ML model used for mobility inference. |
| CMCC | Mobility optimization and UE trajectory are related. We are fine to separate discussion or put UE trajectory in the umbrella of mobility.  In general, mobility will focus on the optimization of mobility decisions and HO parameters via AI-based approach  UE trajectory is to predict the UE position which can be used by mobility function to improve the UE experience and HO performance. |
| LGE | After the detailed use cases are agreed, then we can check the parameters necessary. |
| Ericsson | Our solution description for the mobility optimization is shown below (as proposed in R3-212326)   * The source RAN node can request and obtain from the target RAN node, feedbacks related to the UE, after a successful handover, and obtain an understanding of the UE performance in the target RAN node. Foreseen impact on signaling relates to XnAP and NGAP. * The target RAN node can provide feedback to the source RAN node, e.g. with information related to the load measured at the target RAN node. The source RAN node can use it to refine and improve future handover decisions. Foreseen impact on signaling relates to XnAP and NGAP.   High level signaling flows for the cases described are shown in the figures below. |
| Lenovo, Motorola Mobility | As start point, RAN3 is suggested to focus on the inference data/training data/inference output/measurement result exchange among relevant nodes. Also, the training data provision is relevant to the model update, e.g. when the model update is triggered, the relevant training data shall be provided to the ML training.  If time allows, RAN3 can further the AI model distribution (we consider this is highly dependent on the exact AI algorithm, e.g. if it is a federated learning). |
| Qualcomm | Agree trajectory prediction is a good “toolbox” use case to optimize mobility and also other use cases.  The network can have prediction based on existing UE mobility history (from UE and from gNB), UE’s current location information and radio measurement. |

**Moderator’s summary:**

5 companies think the standard impacts for AI-based use cases should be discussed after the description and details for each use, while a number of companies deem that trajectory/location prediction can be used for mobility optimization, and trajectory/location prediction should be regarded as a separate use case to discuss the corresponding standard impact.

**Proposal 6: AI-based trajectory/location prediction should be regarded as a separate use case to discuss the corresponding standard impact.**

**To be continued:**

**The detailed standard impacts for mobility optimization should continue to discuss at next meeting.**

Companies are invited to provide their views on the input/output for AI-based mobility optimization.

|  |  |
| --- | --- |
| **Company** | **Comments** |
| ZTE | For UE trajectory prediction:  The following data is required as input data for UE trajectory prediction.   * UE historical location information, e.g. Latitude, longitude, altitude * UE moving direction * UE velocity * Radio measurements related to serving cell and neighbouring cells associated with UE location information, e.g., RSRP, RSRQ.   The output of UE trajectory prediction contains the following information.   * UE trajectory prediction (Latitude, longitude, altitude of UE over a future period of time) * The predicted Target cell/NG-RAN node |
| Nokia | Input to ML Mobility Optimization running at a node may be:  -UE Trajectory prediction information (in terms of predicted cells where the UE will connect to) available at RAN.  -Reward (or cost information) from other ML Mobility Optimization algorithms running in neighbouring nodes regarding successful (or unsuccessful) handovers.  Output of ML Mobility Optimization may be a (predicted) Target cell and timing info when a UE is expected to be handed over (BHO, CHO). |
| Deutsche Telekom | We should first focus on detailed use case descriptions (see CB # 46) before going into details of standard impact. |
| Samsung | The input can be   * UE info: trajectory, moving velocity, measurement report * Node info: position, resource status * Neighbor node info: position, resource status, QoS parameters after UE HO (e.g. loss rate, delay, etc.)   The output can be   * predicted HO strategy, inc.:   + predicted HO decision: HO or not HO   + predicted DC activation decision   + predicted HO target node   + predicted HO source node   + predicted HO time   OR the output can be predicted UE trajectory, inc.:   * + predicted UE coordinate/velocity   + predicted camping cells |
| Intel | This is not aligned with use case description in CB #46. |
| vivo | Agree with DT that the standard impacts shall rely on the conclusion of CB #46 |
| NEC | Output of Mobility Optimization is decisions/recommendations for a set of gNBs or gNB-CUs to reconfigure HO related parameters.  Output of UE Position Prediction is time series of predicted coordinates of a UE or of a group of UE.  Output of UE Position Prediction could be used as one of the inputs to Mobility Optimization. |
| Futurewei | Agree with DT and vivo. Standards impact for AI-based use cases should be discussed after the description and details for each use case are agreed (under CB #46). |
| CATT | For prediction:  We agree with the comment by ZTE.  For HO (and potentially CA/DC) decision:  Input should include all the inputs of legacy HO decision mechanism (e.g. Uu measurement result, QoS requirement), as well as predicted UE location, predicted per-UE user plane traffic, and predicted load status of source cell and target cells. |
| Huawei | Similar as above, mobility history info, radio quality measurement info, prediction info (trajectory, load), such info could be used as input; and mobility policy or mobility command as output. |
| LGE | **Input:**   * **Information from CN (the input can be based on the information from AI based CN function):** * UE mobility statistics parameters, e.g., UE location statistics (duration of the time slot) * UE mobility predications, e.g., predicated UE location information in the analytical period * **Information from UE:** * Current and past location statistical information, e.g. GPS, GNSS, cell and the duration information * UE may also have the training model on its locations, thus UE can report the predicated location to RAN   **Output:**  e.g., the predicated target cell IDs or gNB IDs together with the confidence of the predication |
| Ericsson | Possible inputs needed by a mobility optimization AI model:   * Feedback over Xn from target RAN node of the performance of a UE after a successful HO from source to target. Performance may be provided in terms of QoS metrics, QoE, or indication of fulfilment of certain QoA/QoE targets. Target RAN also provides its load metrics at the time of serving the UE * Assistance information from the UE to predict UE trajectory * Information from neighbour nodes about the load or load predictions at neighbour cells   Possible outputs from a mobility optimisation ML model:   * Prediction of UE trajectory * Prediction of UE traffic patterns * Prediction of UE performance at a given potential target cell * Decision on target cell selection |
| Qualcomm | The model can be trained using history data gather from MDT/SON.  In the inference, model input could include:   * UE mobility history information (collected by gNB and received from UE) * UE’s current location * Radio measurement * Ongoing service.   The model output could be either recommended handover targets or arrival probability for each potential targets. |

**Moderator’s summary:**

4 companies think the standard impacts for AI-based use cases should be discussed after the description and details for each use, while a number of companies mentioned that the UE trajectory/location prediction is benefit for mobility optimization. For trajectory/location prediction, UE mobility history information can be the input, and UE trajectory prediction or the predicted target cell can be the output.

**To be continued:**

**For trajectory/location prediction, UE mobility history information can be the input, and UE trajectory prediction or the predicted target cell can be the output.**

### Standard impacts for AI-based Energy saving (if agreed)

Companies are invited to provide their views on the solution and standard impacts for AI-based energy saving.

|  |  |
| --- | --- |
| **Company** | **Comments** |
| ZTE | Energy saving activation decision making may be based on the various information including predicted information. There are many predicted information that can be used by operators to make energy-saving policies, such as energy-saving scenarios classification and traffic load prediction. The more accurate the predicted information is, the better the energy-saving decisions based on the predicted information results will be.  Therefore, AI-based load prediction shall be as a tool-box use case to discuss and we propose the solution to AI-based load prediction. The standard impacts for AI-based load prediction can be seen in Section 3.4.1.  Moreover, energy-saving scenario classification is to cluster similar energy saving scenario through analyzing the various information collected from NG-RAN including traffic tide, coverage status, etc. And energy saving decision can be made for specific energy saving scenario. |
| Nokia | Energy saving use case may have an impact in signaling in OAM interface with RAN and in Xn interface between NG-RAN nodes. |
| Deutsche Telekom | We should first focus on detailed use case descriptions (see CB # 46) before going into details of standard impact. |
| Samsung | There are two types of AI-based energy saving:   * Type 1 ML-assisted energy saving: Energy saving decision is generated by conventional method based on the predicted resource status, where the predicted resource status is generated by ML model. AI functionality is to generate predicted resource status. * Type 2 ML-generated energy saving: Energy saving decision is generated by ML model based on the current/historical resource status. AI functionality is to generate predicted energy saving decision.   So, NG-RAN node is required to support ML model inference, exchange predicted resource status or predicted energy saving decision. |
| Intel | Agree with Nokia. Besides, Xn interface can be used to exchange load prediction information or an indication to avoid deactivating a needed cell of a neighboring gNB (to reduce ping-pong event). |
| vivo | Agree with DT that the standard impacts shall rely on the conclusion of CB #46 |
| NEC | Outputs of Load Balancing Prediction and UE Position Prediction could be used as inputs to Network Energy Saving. |
| Futurewei | Agree with DT and vivo. Standards impact for AI-based use cases should be discussed after the description and details for each use case are agreed (under CB #46). |
| CATT | We believe energy saving can rely on AI-based load prediction, i.e. what ZTE means, and as well as Type1 in Samsung’s comment.  For Type2 in Samsung’s comment, we think we should first identify whether it is really beneficial. |
| Huawei | Here similar as load balancing use case, prediction info, e.g. load prediction, mobility prediction info and mobility history info will be used as input for AI model to work out energy saving policy or direct energy saving action (e.g. switch off cell/carrier). |
| CMCC | We agree with Nokia Energy saving use case may have an impact in signaling in OAM interface with RAN and in Xn interface between NG-RAN nodes.  Both type 1 and type 2 solutions should be considered |
| LGE | After the detailed use cases are agreed, then we can check the parameters necessary. |
| Ericsson | The Energy Saving use case has been described in R3-212315 and can have standardization impacts on the following interfaces:   * Uu:   + Acquisition of information form the UE to help evaluation of energy efficiency configurations at RAN   + Acquisition of information allowing an estimation of traffic/mobility prediction   + Acquisition of information revealing QoE/QoS at the UE * Xn:   + exchange of information describing actions performed for energy saving reasons and possible estimated energy saving gains for such actions   + Exchange of metrics denoting energy efficiency at RAN nodes   A high-level signalling example for the case described is shown in the figure below.  UE  Serving RAN  UE Assistance Information (e.g. Traffic Prediction)  AI deriving RAN configuration for improved EE, e.g. offloading traffic to neighbour RAN  Neighbour RAN  RAN Assistance Information (e.g. Load Measurement/Prediction)  HO Request (Cause = EE improvements, measure of predicted EE improvement)  Accept/Fail depending on EE changes at target RAN |
| Lenovo, Motorola Mobility | As start point, RAN3 is suggested to focus on the inference data/training data/inference output/measurement result exchange among relevant nodes. Also, the training data provision is relevant to the model update, e.g. when the model update is triggered, the relevant training data shall be provided to the ML training.  If time allows, RAN3 can further the AI model distribution (we consider this is highly dependent on the exact AI algorithm, e.g. if it is a federated learning). |
| Qualcomm | Load prediction, traffic prediction and trajectory prediction can be used for network to make on/off decision. We should first define standard to enable the load prediction and traffic prediction. Then, OAM and RAN can take the predictions into consideration in the on/off decision in different level (cell, RAT, BWP, time …). |

**Moderator’s summary:**

5 companies think the standard impacts for AI-based use cases should be discussed after the description and details for each use. A number of companies deem that load prediction and trajectory prediction can be used for energy saving decision, and load prediction/trajectory prediction should be regarded as a separate use case to discuss the corresponding standard impact. Some companies also think energy saving use case may have an impact in signaling in OAM interface with RAN and in Xn interface between NG-RAN nodes.

**Proposal 7: Load prediction/trajectory prediction can be used for energy saving decision, and should be regarded as a separate use case to discuss the corresponding standard impact.**

**To be continued:**

**The detailed standard impacts for energy saving should continue to discuss at next meeting.**

Companies are invited to provide their views on the input/output for AI-based energy saving.

|  |  |
| --- | --- |
| **Company** | **Comments** |
| ZTE | See load prediction standard impact. While for energy-saving scenario classification and system performance evaluation can be performed in OAM, there is no standard impact on RAN side. |
| Nokia | Input to ML Energy Saving running at a RAN node may be UE Trajectory prediction information (in terms of predicted cells where the UE will connect to) available at RAN.  Output of ML Energy Saving may be a predicted switch-on/off decision of a capacity cell. |
| Deutsche Telekom | We should first focus on detailed use case descriptions (see CB # 46) before going into details of standard impact. |
| Samsung | The input for both type 1 and type 2 can be   * node resource status * neighbor resource status   The output can be:   * Type 1: predicted resource status, inc.:   + predicted resource status   + predicted resource special status, inc. sudden-increment, sudden-reduction, high state, low state, etc.   + predicted status corresponding time * Type 2: predicted energy saving, inc.:   + action level, inc. subframe, channel, carrier, device, etc.   + action type, inc. turn on, turn off, etc.   + action source   + action time |
| Intel | By utilizing the predicted load information of the cell and its neighbour cell, the output from ML energy saving can also be the predicted status of switch on/off of a neighbour cell. |
| vivo | Agree with DT that the standard impacts shall rely on the conclusion of CB #46 |
| NEC | Outputs of Load Balancing Prediction and UE Position Prediction could be used as inputs to Network Energy Saving. |
| Futurewei | Agree with DT and vivo. Standards impact for AI-based use cases should be discussed a4fter the description and details for each use case are agreed (under CB #46). |
| Huawei | Similar as above, mobility history info, prediction info (trajectory or load), such info could be used as input; and energy saving policy or direct energy saving action (e.g. switch off cell/carrier) could be as output. |
| Ericsson | Inputs that may be needed:   * Existing load metrics (reported via Resource Status procedures) * Information of RAN node energy consumption * Assistance information from the UE to predict Ue traffic/mobility * Performance feedback from the UE such as QoE or QoS fulfillment indication   Outputs that might be produced:   * Load/Traffic predictions * Actions such as traffic offloading for energy efficiency reasons and a predicted energy efficiency gain associated to the action * Prediction of energy efficiency given current or predicted conditions |
| Qualcomm | Load prediction and trajectory prediction have been discussed in previous questions. Traffic prediction input is FFS. |

**Moderator’s summary:**

4 companies think the standard impacts for AI-based use cases should be discussed after the description and details for each use. And most of companies mentioned load prediction and trajectory prediction can be used for energy saving and should be regarded as a separate use case to discuss the corresponding standard impact.

**To be continued:**

**The detailed standard impacts for energy saving should continue to discuss at next meeting.**

## Others

In [7], it is proposed to standardized signalling procedure for online training, and define model training procedure with MDT/QoE as reference. However, it seems out the scope of RAN3.



Figure : Online training procedure

Companies are invited to provide their comments on this procedure.

|  |  |
| --- | --- |
| **Company** | **Comments** |
| ZTE | This training procedure seems out the scope of RAN3.  For the data collection, some input data for AI can reuse current SON/MDT mechanism, while the new designed signalling should be further studied for some large scale input/output data exchange. |
| Nokia | We do not support to use model training procedure with MDT/QoE as a reference. QoE SI only now completed and WI is about to start, so such would be too preliminary. Also MDT would require more RAN2 involvement and this is a RAN3-led only SI. |
| Deutsche Telekom | The online training – in contrast to initial offline training – should be within RAN3’s scope. But the detailed procedure is dependent on the learning methods and on the use case they are applied to. Therefore, this should be part of use case descriptions covered in CB #46. MDT/QoE can be only an example for this. |
| Samsung | The detailed training process is out of scope. For the specified use case, if requiring online training, the feasibility and necessity need to be shown, and the relevant interface impact of online training is studied based on use case. |
| Intel | Agree with Nokia, seems it’s a more RAN2 related issue. We prefer to postpone in this release SI. |
| vivo | Unnecessary for now.  If the procedure is essential in some use cases, we are ok to further discuss the details. |
| NEC | This could be consider if time allows. |
| Futurewei | Online training procedure and corresponding interface impacts are use case dependent, thus, we recommend this topic being discussed under AI/ML use case CB #46. |
| CATT | Seems out of RAN3’s scope. |
| Huawei | Similar view as above, it is out of RAN3 scope. |
| CMCC | We can largely use SONMDT for data collection, but how it is related to model training is not clear. |
| LGE | Can be studied in the use case session |
| Ericsson | The proposal seems to be out of RAN3 scope, but also it seems to contradict the SID, which clearly states there should be no new nodes or interfaces defined to support AI. The proposal instead seems to rely on the definition of a new Model manager and model aggregator nodes. |
| Qualcomm | First, we believe online training should be standardized on:   * Training configuration * Model transfer, including download and result upload.   Otherwise, inter-vendor interworking is not possible.  In MDT, the OAM configures gNB/UE to perform measurement. Now, if model is managed by OAM, the OAM can configure gNB to perform model training. So, they could be very similar. This is the reason we think MDT/QoE can be used as reference.  For UE/RAN2 impact, I agree it is out of R17 scope. We only consider gNB based model training in R17.  If the online training procedure is standardized, we can also use it for federated learning. |

**Moderator’s summary:**

Some companies deem that the detailed training procedure is out of RAN3 scope, and other companies that the procedure can be studied case by case later.

**Proposal not agreed.**

In [12], it is proposed to involve SA5 for life cycle management of RAN AI/ML, regardless whether RAN AI/ML would be realized in a virtualized way or dedicated way.

Companies are invited to provide their comments on whether to involve SA5 for life cycle management of RAN AI/ML.

|  |  |
| --- | --- |
| **Company** | **Comments** |
| ZTE | No LS is needed at this stage. The work in other groups can be taken into account as usual. |
| Nokia | Sending an LS to SA5 for life cycle management could be OK but at a later stage. We first need to focus on the basic functionality in the RAN. |
| Deutsche Telekom | A LS to SA5 (and SA2) is certainly needed to discuss alignment with AI/ML approaches in OAM domain and LCM for AI models/functions in the RAN (e.g. on initial offline model training, model deployment and updates, etc.). But before sending it, more progress on study details is required. We can consider it at next meeting. |
| Samsung | Support to coordinate with other groups as usual. For LS, prefer to send at a later stage with more progress. |
| Intel | Send LS to SA5 in later stage if needed. |
| vivo | No LS is needed for now.  RAN3 shall reach a consensus on the framework and AI-enabled RAN architecture at first, and then figure out the impacts on OAM after the discussion case by case. After that, an LS is essential to cooperate with SA5. |
| NEC | At this stage it may be too early for LS to SA5. |
| Futurewei | It is premature to send LS to SA5 at this stage. We should wait till more details for RAN AI/ML architecture have been agreed. |
| Huawei | We think an LS is needed, as we could see that discussions were already seen on model management, life cycle management, data accessibility, training procedure etc which actually should be discussed in SA5. |
| CMCC | Premature to send LS to other groups now |
| LGE | Necessary, but later with our progress |
| Ericsson | As stated in the SID, the work should also imply coordination between different groups. However, we are still far from achieving maturity on the topic and an LS to any group is right now premature. |
| Qualcomm | Agree with majority: it is too early to contact SA5.  There are 4 kinds of 5G related AI/ML works ongoing:   * Core network based AI/ML (e.g. SA2 NWDAF) * OAM/SMO based AI/ML (e.g. SA5 MDAS) * RAN based AI/ML (e.g. RAN3 work) * OTT AI/ML carried by 5G (e.g. SA1 work).   It is not necessarily for us to merge our work to or depend our work on OAM based AI/ML. |

**Moderator’s summary:**

Majority of companies consider that it’s too early to send LS to SA5 at the stage of immature progress.

**The LS to SA5 is not agreed.**

# Conclusion, Recommendations [if needed]

If needed

# References

1. R3-211684, TP to TR 37.817 Solutions and standard impacts (NEC)
2. R3-211858, Standards impact for identified use cases (CATT)
3. R3-211970, Discussion on Standard Impact for RAN Intelligence (Samsung)
4. R3-212034, Initial analyse on the interface impact with AI-based RAN architecture (ZTE Corporation, China Unicom)
5. R3-212371, AI/ML Architecture Discussion (Nokia, Nokia Shanghai Bell)
6. R3-212374, General Principles of ML Functionality in RAN (Nokia, Nokia Shanghai Bell)
7. R3-211755, Model training procedure (Qualcomm Incorporated)
8. R3-212505, Solutions for AI-based load balancing (CMCC)
9. R3-212465, Discussion on Standard Impact for RAN Intelligence (Mobility Management) (Samsung)
10. R3-212546, Support of AI enabled Mobility for NG-RAN and EN-DC (LG Electronics) (Mobility Optimization)
11. R3-212180, Discussion on standard impact to support AI functionality (Lenovo, Motorola Mobility)
12. R3-212525, Discussion on the correlation with other groups (Huawei)