**3GPP TSG-RAN WG2 #128 *R2-24XXXXX***

**Orlando, Florida, USA Nov. 18th – 22th, 2024**

Agenda Item: 8.3.2.1

Source: OPPO(rapporteur)

Title: Draft Summary of [POST127bis][022][AI mobility] Simulation Assumption of measurement event/RLF prediction and SLS (OPPO)

Document for: Discussion, Decision

# Introduction

This draft intends to kick off following email discussion:

* [POST127bis][022][AI mobility] Simulation Assumption of measurement event/RLF prediction and SLS (OPPO)

      Intended outcome: Agreeable Simulation assumptions of measurement event/RLF prediction and SLS for HO performance. Open issues related to simulation assumptions e.g. definition, methodology, metrics etc. can be also covered.

      Deadline:  Nov. 8th

Note this email is the extension of existing email discussion as following:

* [POST127][029][AI mobility] RLF Simulation Assumption (Oppo)

Intended outcome: Agreeable simulation assumption

Deadline: extra long

Although@RAN2#127meeting, RAN2 concluded on definition and metric aspect for RLF, but rapporteur believes some further clarification on definition, methodology and metrics of RLF prediction are necessary in order to discuss simulation assumptions thoroughly.

In RAN2#127bis meeting, RAN2 made progress on measurement event w.r.t. to definition, methodology, metrics and simulation assumption. As RAN2 aims to conclude all relevant open issues in RAN2#128 meeting in November and RAN2 agreed to cover measurement event prediction in existing email threshold[POST127][029], open issues of this use case is covered now. The relevant agreement is:

* Aim to narrow down options and values. Will continue this discussion over email discussion

In the same meeting, RAN2 agreed to cover SLS simulation assumptions issues also in this email thread. The related agreements are:

**Agreements**

1. The System level performance (e.g. HO performance) evaluation is optional (i.e. companies can bring results if they chose).
   1. System level performance for measurement event prediction can be prioritized by companies if they chose to do it.

1. RAN2 will prioritize discussions on intermediate KPI discussion before discussing system level performance for the corresponding use case.

2: Discussion on what (type of information)/how generalization study can take place in Nov. meeting

3: The SLS simulation assumption discussion is covered in the post#127bis email discussion by assuming:

 The simulation assumptions agreed for measurement event prediction and RLF prediction is taken as baseline for SLS in principle

 The HO model in 36.839 is taken as baseline

 The HO performance will be HOF and number of HO only and definition in 36.839 is taken as baseline

 The baseline of HO performance is R15 legacy measurement and HO procedure

# Discussion

## Measurement event prediction

### Definition

RAN2 agreed@127bis meeting following input and output for indirect prediction measurement event:

1. For indirect measurement event prediction, the intermediate output (i.e., the output of RRM prediction model) is RSRP of serving/neighboring cells. The final output is the expected occurrence time of a certain measurement event (ex. event A3).

Here is one example algorithm structure cited from [1]:



Figure 2.1.1-1 algorithm structure of indirect measurement prediction

RAN2 agreed to focus on FR2 temporal domain case A and optional FR1 temporal domain case B for measurement event prediction. For case B, the post processing unit will collect both actual measurement results and predicted measurement results. Measurement event can be deduced based on collected measurement results by following specified rules including assess entry condition and event occurrence. At any point of time there could be also predicted measurement results in PW. So naturally both historical actual/predicted (and skipped) measurement results and future predicted measurement results can be used for assessment purpose. For temporal domain case A, it is possible that only predicted measurement results are used assuming PW length is longer or equal to TTT length. But when PW length is shorter than TTT length then both historical actual measurement results and future predicted measurement results are both involved. As for FR1 inter-frequency prediction it is obvious that only current and historical predicted measurement results are involved. Having said that, here is recommended definition of indirect measurement event prediction for temporal domain case A, case B and frequency domain scenario:

**Indirect measurement event prediction for temporal domain case A:**

In indirect measurement event prediction, future measurement result(s) is predicted by a RRM measurement prediction model in temporal domain at first, based on which and optional historical actual measurement result(s), a measurement event at one future time instance is derived without involvement of further AI/ML model.

**Indirect measurement event prediction for temporal domain case B:**

In indirect measurement event prediction, measurement result(s) is predicted by a RRM measurement prediction model in temporal domain at first, based on which and historical actual measurement result(s) a measurement event at one time instance is derived without involvement of further AI/ML model.

**Indirect measurement event prediction for frequency domain:**

In indirect measurement event prediction, measurement result(s) is predicted by a RRM measurement prediction model in frequency domain at first, based on which and historical actual measurement result(s) a measurement event at one time instance is derived without involvement of further AI/ML model.

*Note: The difference compared to case A is marked for case B and frequency domain by change mark.*

**Question 1: Do you agree those 3 recommended definitions of indirect measurement event prediction?**

|  |  |  |
| --- | --- | --- |
| Company | Opinion: Yes or No | Comments |
| vivo | Yes with comments | For Case A, no need to consider the historical actual measurement result(s) for the following reasons:   * The goal of Case A is to improve handover performance, thus the time for event triggering needs to be sufficiently advanced so that the UE can receive the appropriate handover command in a timely manner. * The prediction accuracy of RSRP is expected to be good enough, and the measurement event is expected to be triggered when the entering conditions of the event are not yet met/just met, rather than waiting for a duration of actual measurement. * Historical measurements only need to be considered when the PW length is shorter than TTT, but currently the PW is longer than TTT in the simulation assumption.   For temporal domain case B, historical actual measurement results(s) is optional. For example, when TTT is shorter than the measurement period, measurement event can be determined as triggered only with one predicted measurement result.  For frequency domain, no actual measurement on the frequency for prediction. Thus the historical actual measurement result(s) should be removed as well.  The Definition of indirect measurement event for spatial domain should be added, e.g.,  In indirect measurement event prediction, measurement result(s) is predicted by a RRM measurement prediction model in spatial domain at first, based on which a measurement event at one time instance is derived without involvement of further AI/ML model. |
| Xiaomi | Yes for case A, B  Comments on frequency domain | In frequency prediction, assume AI predict the measurement results on FRa based on actual measurement results on FRb.  Here, we are talking about the event predition on FRa. If this is correct understanding, there is no actual measurent result on the the FRa. So, the event is evaluated based on pure predicted results. Actual measurement on FRb is not used to evaluate event on FRa.  In addition, we suggest to remove the ‘at first’ in all definitions, which is confusing. Because actual measurement occurs before prediction. |

For direct measurement event prediction, RRM measurement model can’t be reused because the output of the model is an event. But the input of RRM measurement model can be reused since nothing new is proposed by company @127bis meeting. As for the output it is not clear how to interpret the time window. For temporal domain case A, on one hand it would be desirable that the window is in future as illustrated in Figure 2.1.1-2:



Figure 2.1.1-2 interpretation 1 of event occurrence window

The problem for such interpretation is that it could be difficult for model to do so. The reason to have probability is because the model can’t make sure whether there is an event in the window (between t1 and t2 in Figure 2.1.1-2) for 100%. But it also implies the difficulty for the model to predict there is no event in the time window behind (between t0 and t1 in Figure 2.1.1-2) for 100%. So, another interpretation is that the start of the window should be current time instance i.e. t0 as illustrated in Figure 2.1.1-3:



Figure 2.1.1-3 interpretation 2 of event occurrence window

The problem for interpretation 2 is that the output of the model is not so valuable for network to take action to enhance HO performance.

The problem is for the study targeting 2nd goal i.e. to enhancement HO performance. If the study is targeting 1st goal e.g. for FR1 temporal domain case B, it is not clear which interpretation is applied considering some of measurement results are skipped. Since RAN2 agreed to focus on FR2 temporal domain case A for measurement prediction, rapporteur suggestion to focus on indirect prediction for FR1 temporal domain case B.

**Question 2: Do you agree that the input of model for RRM measurement prediction is reused for direct measurement event prediction?**

|  |  |  |
| --- | --- | --- |
| Company | Opinion: Yes or No | Comments |
| vivo | Yes with comments | Similar to RRM measurement prediction, companies are suggested to report the input in the simulation results. |
| Xiaomi | Yes as baseline | As baseline. The event prediction is done at UE side, so maybe more inputs can be considered, e.g. Ue location. |

**Question 3: If the study is targeting 2nd goal in this SI, how to interpret the event occurrence window? This question is only applied for temporal domain.**

|  |  |  |
| --- | --- | --- |
| Company | Opinion: interpretation 1, interpretation 2 or others | Comments |
| vivo | Interpretation 2 | If the interval between T0 and T1 is too long, the network does not need to handover the UE to another cell too early. Before receiving the HO command, the event prediction results within the time window may change, and the UE needs to report the updated prediction result, causing unnecessary signaling overhead.  Besides, the prediction accuracy of interpretation 2 is expected to be better as the time window is closer to the OW.  In addition, the time windows of direct and indirect approaches should be consistent. Since the starting point of PW for the indirect approach is the current time, the start of the window for direct prediction should be the current time instance as well. |
| Xiaomi | Combination of two interpretations | We understand the direct prediction can predicit the event in multiple windows. The window(s) are consecutive and the first window starts from t0. Following is an example    For a specific window, the start can be at or after t0. |

**Question 4: If the study is targeting 1st goal in this SI i.e.FR1 temporal domain case B in the context, can we focus on indirect prediction only?**

|  |  |  |
| --- | --- | --- |
| Company | Opinion: Yes or No | Comments |
| vivo | Yes |  |
| Xiaomi | Yes |  |

Based on RAN2 agreement and above discussion, we can have following definition for direct measurement event prediction:

**In direct measurement event prediction, a measurement event within a time window is predicted with possibility x% directly, where 0<x<=100, based on same input of model for corresponding RRM measurement use case**

Once question 2/3/4 are answered, then we can improve the wording of the definition.

**Question 5: Do you agree the recommended definition of indirect measurement event prediction as baseline for further improvement?**

|  |  |  |
| --- | --- | --- |
| Company | Opinion: Yes or No | Comments |
| vivo | Yes with comments | The phrase ‘**based on same input of model for corresponding RRM measurement use case**’ is misleading as there is no corresponding RRM measurement prediction for the direct approach. Therefore, suggest refining as:  **In direct measurement event prediction, a measurement event within a time window is predicted with possibility x% directly, where 0<x<=100, based on the same input of ~~model for corresponding~~ RRM measurement prediction use case** |
| Xiaomi | Yes with comments | We don’t need to restric the input. The event prediction is done at UE side, so maybe more inputs can be considered, e.g. Ue location. |

### Metrics

F1 score is a classical performance metrics for classification model. In general, the model intends to predict an event, which may occur in baseline case. And depending on whether an event is predicted in AI case i.e. by the model and/or occur in baseline case, the combination of AI case and baseline case could be illustrated in Table 2.1.2-1:

|  |  |  |
| --- | --- | --- |
| true event\predicted event | Negative | Positive |
| Negative | n0 | n1 |
| Positive | n2 | n3 |

Table 2.1.2-1 counters to calculate F1 score

Here are general interpretations of those counters in the table 2.1.2-1:

Counter n0: refers to the case where no event is predicted and occur in baseline case. This is not used for F1 score calculation.

Counter n1: refers to the case where an event is predicted while it doesn’t occur in baseline case i.e. model’s prediction is kind false alarm

Counter n2: refers to the case where no event is predicted while it occurs in baseline case i.e., an real event is missed by the model

Counter n3: refers to the case where an event is predicted while it does occur in baseline case i.e. model predicts an event correctly.

The normalized metrics are:

Precision = n3/(n1+n3) Formula\_1

Recall =n3/(n2+n3) Formula\_2

F1 score = 2\*Precision\*Recall/(Precision + Recall) Formula\_3

RAN2 agreed that F1 score as well as component counters (with different terms) for measurement event prediction and RLF event prediction. Here is relationship between counters in table 2.1.2-1 and agreed metrics for measurement event prediction and RLF event prediction:

|  |  |  |
| --- | --- | --- |
| Counter | Metrics of measurement event prediction | Metrics of RLF prediction |
| n1 | False event prediction | False event prediction |
| n2 | Missed event prediction | Missed event prediction |
| n3 | True event prediction | True event prediction |

Table 2.1.2-2 mapping between component counters of F1 score and agreed intermediate metrics

For both indirect prediction and direct prediction, the meaning of “true event prediction” should be clarified in order define the rest two counters. For indirect prediction the agreed metric i.e. “time difference of true time event and predicted time event”(let’s call it ETD i.e. event timing difference) can be used to define the meaning of “true event prediction”.



Figure 2.1.2-1

In Figure 2.1.2-1, when an event is predicted and there is a real event occurs around this predicted event whose range is [0, max ETD] then it is a “true event prediction”. Note the relationship between predicted event and real event is symmetry. In Figure 2.1.2-1, If an event is predicted but there is no real event falls within the yellow window, then it is a “false event prediction”. If a real event occurs but there is no predicted event falls within the yellow window, then it is a “missed event prediction”. Here is recommended definition of these 3 counters:

**Counter n3: it increases by 1 when a real event occurs around a predicted event with ETD, whose range is [0, maximum ETD] or vice versa**

**Counter n1: it increases by 1 when no real event occurs around a predicted event with ETD, whose range is [0, maximum ETD]**

**Counter n2: it increases by 1 when no event is predicted around a real event with ETD, whose range is [0, maximum ETD]**

**Question 6: Do you agree above definitions of true event prediction, false event detection and missed event detection for indirect measurement event prediction?**

|  |  |  |
| --- | --- | --- |
| Company | Opinion: Yes or No | Comments |
| vivo | Yes |  |
| Xiaomi | Yes |  |

For direct prediction the story is bit different because of the output of the model is probability of event occurrence within a time window. One proposal from [2] is to check whether a true event falls within the time window of the predicted event. If it does, it is true event prediction otherwise it is kind of false alarm. On the contrary if a real event occurs but it doesn’t fall within time window of any predicted event, then this event is missed by the model. Based on this here are recommended definitions:

**Counter n3’: it increases by 1 when a real event occurs within the occurrence window of predicted event whose possibility is higher than a predefined threshold**

**Counter n1’: it increases by 1 when no real event occurs within the occurrence window of predicted event whose possibility is higher than a predefined threshold**

**Counter n2’: it increases by 1 when a real event occurs, but it doesn’t fall in the occurrence window of any predicted event whose possibility is higher than a predefined threshold**

**Question 7: Do you agree above definitions of true event prediction, false event detection and missed event detection for direct measurement event prediction?**

|  |  |  |
| --- | --- | --- |
| Company | Opinion: Yes or No | Comments |
| vivo | Yes |  |
| Xiaomi | Yes |  |

The F1 score is then defined based on Formula\_1, Formula\_2 and Formula\_3. Naturally F1 score is a comprehensive intermediate metric for measurement event prediction. It reflects the performance of the model as whole. On the other hand, Foluma\_1 and Formula\_2 also reflect how truly and precisely the model can predict an event. The question is whether these two metrics can be also used as intermediate metric?

**Question 8: Apart from F1 score, can we also use the metrics defined in Formula\_1 and Formula\_2 as intermediate metrics for measurement event prediction?**

|  |  |  |
| --- | --- | --- |
| Company | Opinion: Yes or No | Comments |
| vivo | Yes with comments | The KPIs related to Formula 1/2 were agreed at the last meeting, i.e.,   * for indirect: F1 score. the following can be reported: RSRP difference, missed event detection, false event detection. FFS how to define F1 score.   These two intermediate KPIs can reflect the impact on mobility: With the false detection, too-early HO or HO to wrong cell may occur; Upon miss detection, too-late HO may happen.  Besides, companies are encouraged to provide the confusion matrix as in Table 2.1.2-1, so that any metrics can be derived when needed. |
| Xiaomi | No strong view |  |

### Simulation assumption

@RAN2#127bis RAN2 agreed:

* Measurement event prediction simulations will at least focus on intra-frequency FR2, case A, and second study goal (i.e. HO KPI improvement). FFS what is KPI.
* Companies can bring simulation results for intra-frequency measurement reduction for FR1 and report what they are doing. Focus on temporal case B.
* Companies will prioritize simulations on indirect method. Companies can bring simulations on direct method and should report what method is being used.
* Measurement event prediction results are expected in RAN2#129

And for simulation assumptions, RAN2 agreed a principle:

* *The Simulation assumption of RRM measurement prediction can be reused unless otherwise specified.*

Since both FR1 intra-frequency temporal domain case B and FR2 intra-frequency temporal domain case A are on the table, the simulation of both case A and case B will be discussed.

For case A, following agreed principle above, Table 5.1-1 in the text proposal [3] is taken as baseline. On top of that RAN2 also agreed to narrow down some key parameters when discussing [4]. For case A, the key parameter to impact the general trends are UE speed and OW/PW length. The event parameters are A3 event offset and TTT when discussion [5]. Here is the additional parameters and recommended value from rapporteur:

|  |  |
| --- | --- |
| Parameters | Recommended value |
| A3 event offset (db) | 2 |
| TTT (ms) | 160 |
| UE speed (km/h) | 90 |
| OW length (ms) | 800 |
| PW length (ms) | 400 |
| Max ETD (ms, note1) | 80 |
| Event occurrence Window Length (ms, note 2) | 160 |
| Probability threshold (%, note 2) | 50% |

Table 2.1.3-1 additional parameters for case A

*Note1: parameters for indirect prediction*

*Note2: parameters for direct prediction*

**Question 9: For FR2 intra-frequency temporal domain case A, on the top of the parameters in table 5.1-1 in [3], can you agree with the recommended parameters and corresponding values in table 2.1.3-1? If you think any parameter is missed, please provide recommended value.**

|  |  |  |
| --- | --- | --- |
| Company | Opinion: Yes or No | Comments |
| vivo | No for TTT and OW,  Yes for others | The TTT should be set with two typical values rather than a single one, e.g., a short value (e.g., 80ms) and a long value (e.g., 320ms), so that we can see the baseline performance of legacy HO with different TTT settings.  For the OW, in our previous simulation, when the OW and PW are the same length, the accuracy is almost the best. Therefore, no need to align the OW length, and companies can select the optimized length and report the value. |
| Xiaomi | No for OW,PW | As the TTT is only 160ms, we don’t see the need for UE to predict event fulfillment in future 400ms. The gain of event prediction comes from saving the HO preparation delay. So, shorter PW value should be allowed.  Also, as the event prediction is done at UE, it can be up to UE to decide the OW length, which may has no impact to spec.  Companies can report the used OW,PW length. |

Similarly, for FR1 intra-frequency temporal domain case B, on top of the parameters in table 5.1-1 in [3], the parameters in following table 2.1.3-2 can be also discussed:

|  |  |
| --- | --- |
| Parameters | Recommended value |
| A3 event offset (db) | 2 |
| TTT (ms) | 160 |
| UE speed (km/h) | 30 |
| OW length (ms) | 400 |
| PW length (ms) | 200 |
| Max ETD (ms, note1) | 40 |

Table 2.1.3-2 additional parameters for case B

**Question 10: For FR1 intra-frequency temporal domain case B, on the top of the parameters in table 5.1-1 in [3], can you agree with the recommended parameters and corresponding values in table 2.1.3-2? If you think any parameter is missed, please provide recommended value also.**

|  |  |  |
| --- | --- | --- |
| Company | Opinion: Yes or No | Comments |
| vivo | No for TTT and OW，  MRRT should be added,  Yes for others | Similar to Q9, the TTT should be set with two typical values rather than a single one.  Similar to Q9, the OW length does not need to be aligned.  For case B, the MRRT should be aligned (e.g., 1/2). |
| Xiaomi | Yes |  |

RAN2 agreed that sub-case2 will be focused for measurement event prediction. During simulation the input L3 filtered RSRP result is achieved usually based on L1 beam results of dataset. In case some measurements are skipped, it is most likely that L1 beam measurement results are skipped but not L3 filtered RSRP result because otherwise almost no measurement can be saved especially for UE sided model. Then the issue is how to do the filtering for the input L3 RSRP results. Contribution [6] provide 3 options:

**Observation 1: For temporal domain prediction case B, the following filtering scheme can be considered:**

* **Filtering option 1: L3 filtering is based on the L1 filtered result and the last L3 filtered cell result;**
* **Filtering option 2: L3 filtering is based on the L1 filtered result if the last L3 filtered result is from prediction;**
* **Filtering option 3: L3 filtering is based on the L1 filtered result and the L3 filtered result from last actual measurement.**

Take following as example:



Figure 2.1.3-1: Temporal domain case B example 2

The 3 options also mean:

* Filtering option 1: L3 cell result at T3 is filtered based on the L1 filtered result at T3 and L3 filtered cell result at T2;
* Filtering option 2: L3 cell result at T3 is filtered based on the L1 filtered result at T3; in other words, not consider old L3 filtered result.
* Filtering option 3: L3 cell result at T3 is filtered based on the L1 filtered result at T3 and L3 filtered cell result at T1 (i.e. the old L3 filtered cell result from last actual measurement);

Rapporteur think it is important to clarify how company do the filtering for input L3 RSRP. Since the email discussion is rather short and filtering algorithm itself is usually an implementation issue, it is recommended to report the options for input L3 RSRP result for sub-case 2.

**Question 11: For FR1 intra-frequency temporal domain case B and sub-case 2, do you agree to report the filtering option for input L3 RSRP result?**

|  |  |  |
| --- | --- | --- |
| Company | Opinion: Yes or No | Comments |
| vivo | Option 1 only | Option 1 is aligned with the legacy L3 filtering process, i.e., both the latest value and the previous value are considered to reduce the impact of rapid RSRP changes.  Option 2 completely disregards L3 filtering, which will lead to significant RSRP changes. Besides, it may cause inconsistency between the test set and the training set.  Option 3 skips certain time points for filtering, which may also cause inconsistency between the test set and the training set as it ignores the RSRP change trend at those time points. |
| Xiaomi | Option 1 | RAN2 has discusse this issue for RRM in RAN2 127. Following is agreed,  In the definition of 3 RRM sub-cases, all cell level measurement result(s) refers to L3 filtered cell level measurement  Now, seems option 2 and 3 are new L3 filtering method and not aligned with RRM prediction.  We wonder why different assumptions is made to event prediction. |

## RLF event prediction

### Definition

Since RAN2 agreed @RAN2#127 meeting that SINR is the measurement quantity, so it is clear that L1 SINR results of RLF detection reference signal are the input of the model for both direct and indirect RLF prediction.

For indirect RLF prediction, the output of the model is L1 SINR of RLF detection RS in the future. And then RLF event is derived. In order to predict future L1 SINR, it makes sense to follow methodology as temporal domain case A.

The assumed T310 is 1000ms in 36.839. If prediction window is equal or longer than T310, then only predicted L1 SINR will be used for derivation. But if prediction window is shorter than T310, then actual measured L1 SINR will be also involved in the assessment procedure. Based on this analysis and RAN2’s agreement @ RAN2#127 meeting, here is recommended definition of indirect RLF:

**Indirect RLF prediction:**

**In indirect RLF event prediction, future L1 SINR measurement result(s) of serving cell is predicted by a RRM measurement prediction model in temporal domain at first, based on which and optional historical actual L1 SINR measurement result(s) a RLF event at one future time instance is derived without further AI/ML model.**

**Question 12: Do you agree recommended definition of indirect RLF prediction?**

|  |  |  |
| --- | --- | --- |
| Company | Opinion: Yes or No | Comments |
| vivo | Yes with comments | The L1 SINR should be further clarified as L1-filtered SINR. |
| Xiaomi | Yes | The key point is the RLF modelling is the same, i.e. consecutive OOC during T310 in AI and non-AI. In AI, the OOC can be based on prediction and measurement. If the prediction window is longer than T310, UE can predict RLF before T310 is triggered. Otherwise, UE can only predict RLF after T310 is triggered. |

For indirect RLF prediction, the **Question 3** about occurrence window options as illustrated in Figure 2.1.1-2 and Figure 2.1.1-3 for measurement event prediction is also valid. So similar questions are also raised here. Note the study goal of RLF prediction so far focus on goal 2:

**Question 13: For direct RLF prediction, how to interpret the event occurrence window?**

|  |  |  |
| --- | --- | --- |
| Company | Opinion: Yes or No | Comments |
| vivo | Interpretation 2 | Similar comments to Q3. |
| Xiaomi | Combination of two | We understand UE can predict RLF in multiple windows in direct prediction. The window(s) are consecutive and the first window starts from t0. Following is an example    For a specific window, the start can be at or after t0. |

Based on RAN2 agreement, we can have an explicit definition of direct RLF prediction:

**In direct RLF event prediction, a RLF event within a occurrence window is predicted with possibility x% directly, where 0<x<=100, based on historical actual L1 SINR measurement results**

Once question 13 is answered, the we can improve the definition of direct RLF event prediction.

**Question 14: Do you agree the recommended definition for direct RLF prediction as baseline?**

|  |  |  |
| --- | --- | --- |
| Company | Opinion: Yes or No | Comments |
| vivo | Yes |  |
| Xiaomi | Yes |  |

### Metrics

It is rapporteur’s understand that the metrics for measurement event prediction can be reused for RLF prediction.

**Question 15: Do you agree metrics concluded in section 2.1.2 are reused for direct and indirect RLF event prediction?**

|  |  |  |
| --- | --- | --- |
| Company | Opinion: Yes or No | Comments |
| vivo | Yes | Similar comments to Q8, companies are encouraged to provide the confusion matrix. |
| Xiaomi | Yes with comments | However, we want to point out the impact to NW of miss RLF detection and false RLF detection is not the same. If AI miss RLF detection, there is no harm, since the consequence is the same as legacy, i.e. RLF. If AI false predict RLF, it may harm the NW. Because NW may decide to handover the UE or UE would trigger RRC reestablishment due to false prediction, which would result in additional interruption. So, we should prioritize the evaluation of false RLF prediction. |

### Simulation assumption

In RAN2#127 it was confirmed that simulation assumption for FR2 RRM measurement prediction is reused as baseline. In RAN2#127bis RAN2 endorse a text proposal [3] which also covers the simulation assumptions for both FR1 and FR2. The additional part is the TX/RX beam number. Considering RLF parameters could be reused for SLS, where both FR1 and FR2 case will be covered, we need confirm a baseline assumption for FR1 and FR2 for RLF prediction.

**Q16: Do you agree table 5.1-1 in [3] is taken as baseline simulation assumption for RLF prediction for both FR1 and FR2?**

|  |  |  |
| --- | --- | --- |
| Company | Opinion: Yes or No | Comments |
| vivo | Yes |  |
| Xiaomi | Yes |  |

Apart from baseline simulation assumptions, more parameters are proposed at RAN2@127meeting. Among the contributions, there are 5 kinds of simulation parameters:

Case 1: parameters related to channel model e.g. channel blockage [1] [12]

Case 2: parameters related to RLF procedure [1] [9] [12] [8][7]

Case 3: parameters related to HO procedure [8]

Case 4: parameters related to UE’s characteristics e.g. speed, trajectory etc. [10]

Case 5: Interference related parameters [1]

Parameters in case 3 can be saved because RAN2 agreed to simulate RLF without HO procedure. Parameters in case 4 can be also saved because company can report UE’s characteristics parameters, which is already in the template of the spreadsheet after [POST127bis][016][AI Mob] Simulation results (Mediatek). The intention for parameters in case 1 is to trigger RLF event easier. But now RAN2 agreed that UE will not change serving cell after it is dropped, triggering RLF event is not difficult any more. Considering company would like to reuse existing data set as much as possible, no change of such parameter for channel model is preferred.

**Observation 2: parameters in case 2, case 3 and case 4 can be saved.**

For parameters in case 2, some companies [9][11][13] propose parameters to enable easier RLF triggering e.g. a short T310 timer. Again, since no HO procedure will be simulated, such change is not necessary from this perspective. However likely same criteria and parameter will be used for SLS, so a shorter T310 timer is necessary. Having said that, table 2.2.2-1 summarize the additional parameters for RLF prediction:

|  |  |
| --- | --- |
| Parameter | Value |
| Qin threshold | -6db |
| Qout threshold | -8db |
| Sample rate (TIndication\_interval) | 20ms(FR2)/40ms(FR1) |
| Qin evaluation period | 100ms |
| Qout evaluation period | 200ms |
| T310 | 200ms |
| N310 | 1 |
| N311 | 1 |
| Max ETD (ms, note1) | 20ms(FR2)/40ms(FR1) |
| Event occurrence Window Length (ms, note 2) | 40ms(FR2)/80ms(FR1) |
| Probability threshold (%, note 2) | 50% |

Table 2.2.2-1 Additional RLF parameters for FR2 and FR1

*Note1: parameters for indirect prediction*

*Note2: parameters for direct prediction*

You may notice that the recommended sampling rate is 20ms instead of 80ms for FR2. In NR, when DRX is not used, TIndication\_interval is max(10ms, TRLM-RS,M) according section 8.1.6 of 38.133:



If taking TSSB as example, sample rate could be 10,20,40,80 ms. In [POST127][030][AI mobility] RRM simulation assumptions (OPPO), we agreed to change sampling period of FR2 from 20ms to 80ms. If 10ms is used for sample rate of RLF prediction on FR2, likely we can’t reuse existing data set as much as possible. Because of this, rapporteur recommend to use 20ms for FR2. For FR1 40ms can be reused also for RLF prediction.

**Question 16: Do you agree the additional RLF parameters in table 2.2.2-1?**

|  |  |  |
| --- | --- | --- |
| Company | Opinion: Yes or No | Comments |
| vivo | Yes |  |
| Xiaomi | Yes |  |

The difficult part is how to simulate interference (case 5 parameters in the context). RAN2 agreed not to simulate traffic but also keep some FFS i.e. “full buffer and assumption that all the cells are fully loaded. We will not simulate traffic”. Before we dive deeper into the discussion, let’s first look at the definitions of those concepts.

Full buffer model is a kind of traffic model that is widely used by 3GPP. The model assumes that UEs always have data to transmit or receive, and their buffers are always full. This reflects a scenario where the system is operating at full capacity, and it is used to study the system's performance under saturated conditions. Compared to other traffic model (e.g., FTP and VoIP) which need to stimulate the stochasticity of data arrival and storage, full buffer model is much simpler. It can be seen as a simple assumption applied and does not need additional work for simulation.

“All the cells are full loaded” refers to a case that all cells are using the same wireless resources (i.e., resource block) for DL transmission. A UE will receive interference from all neighbour cells. In this way, interference is more common and RLF will be easier to trigger.

If we adopt that “full buffer and assumption that all the cells are fully loaded”, we do not need to model a scheduler to model the wireless resource allocation procedure, which is rather complicated. The simulation workload can be saved.

Based on the above analysis, it seems we can directly remove the FFS in the agreements.

**Question 17: Do you agree to use the full buffer assumption and the scenario that all the cells are fully loaded for interference modelling and no resource scheduler is needed?**

|  |  |  |
| --- | --- | --- |
| Company | Opinion: Yes or No | Comments |
| vivo | Yes |  |
| Xiaomi | Yes |  |



Figure 2.3-1: ways to generate interference

From figure 2.3-1(a), we can see that a UE in cell 1 get more interference than a UE within other cells if only those 7 cells are considered. It is a kind of artificial boundary effect, which is not representative of how radio signals behave in practice. Therefore, in simulation, we often use wrap-around to smooth transition of signal strength across the entire simulation area. As we consider 2-tier deployment, the majority of interference comes from the surrounding six cells. An example of wrap-around would be that UEs in cell 6 receive interference from cells 1, 5, 7, 4a, 2b, and 3b, as shown in figure 2.3-1 (b). There is also another way to generate interference as shown in figure 2.3-1(c), where a 3-tier scenario is modelled to get the interference surrounding cell 6. It has the same effect as wrap-around and also use the surrounding six sites for interference analysis. Besides the adjacent 6 sites, interference can also be seen from two co-site cells that are not the serving cell (as we consider 3 sectors/cells per site).

**Question 18: Do you agree that interference in simulation comes from co-site cells and surrounding 6 sites of serving cell, i.e., interference comes from 20 cells?**

|  |  |  |
| --- | --- | --- |
| Company | Opinion: Yes or No | Comments |
| vivo | Yes |  |
| Xiaomi | Yes |  |

We have agreed in RAN2#126 that FR2 study will be prioritized for RLF prediction. It means beams must be considered. From 38.213, RLM using the associated SS/PBCH block when the associated SS/PBCH block index is provided by RadioLinkMonitoringRS. That means RLM-RS can come from beam(s) that are not serving. Given that different companies may use different number of DL beams and different ways to index beams, even if we give a fixed beam pattern for RLM, the SINR may deviate much among companies. Therefore, a simple way is to use the serving beam for interference analysis.

**Question 19: Which option would you prefer to determine the serving signal of RLM:**

* Option 1: a fixed beam pattern that may differ from the serving beam, e.g., beams indexed with {0,2,4,6}.
* Option 2: serving beam

If you prefer option 1, please also provide the preferred number of beams (RLM-RS) and the way to choose them. If you have any other options, please list them in the comments.

|  |  |  |
| --- | --- | --- |
| Company | Option: 1 or 2 | Comments |
| vivo | Option 2 with comments | The serving beam should be clarified as the best beam of the serving cell to ensure that companies have the same understanding of the serving beam. |
| Xiaomi | 2 | We think it’s more practical to consider serving beam. But we don’t need to simulate BM procedure. Maybe the strongest beam can be assumed as serving beam. |

If answer to Q8 is yes i.e. no scheduler is modelled, we also need to discuss how to determine the interference from each cell. From the previous discussion, there are three ways on the table.

* Option 1: the interference comes from the beam with maximum RSRP of neighbor cells
* Option 2: the interference comes from fixed beam(s) of neighbor cells.
* Option 3: the interference comes from randomly chosen beam(s) of neighbor cells.

Option 1 is easier to generate a small SINR but does not strictly follow the current specification. While option 2 aligns more with existing protocols, the modeling of it faces the same issue as we discussed in **Q9**. It would be hard to compare companies’ results as we may use different number of DL beams and indexing method. Option 3 can model the randomness of interference but can lead to possibly large fluctuation of SINRs, making AI hard to learn and predict.

**Question 20: Which option would you prefer to determine interference signals of neighboring cells:**

* Option 1: the interference comes from the beam with maximum RSRP of neighbor cells
* Option 2: the interference comes from fixed beam pattern of neighbor cells.
* Option 3: the interference comes from randomly chosen beam(s) of neighbor cells.

If you prefer option 2 or 3, please provide the preferred number of beams of each cell used for interference. If you pick option 2, please also provide ways to choose the beams. If you have any other options, please list them in the comments.

|  |  |  |
| --- | --- | --- |
| Company | Opinion: 1, 2, or 3 | Comments |
| vivo | Option 2 | Opt 1 is not a reasonable network scheduling implementation and Opt 3 will make the SINR unpredicted because the interface will be unpredicted with random interference beams. Therefore, Opt 2 should be adopted. |
| Xiaomi | Option 1 |  |

**Question 21: Do you believe there are any other important factors we should consider for simulating interference? If so, please provide them below.**

|  |  |
| --- | --- |
| Company | Comments |
| vivo | For option 2 in Q20, suggest clarifying that the “fixed beam pattern of neighbor cells” is the same for different random seeds, since it will impact the RLF prediction performance. |
| Xiaomi | We understand the RLF may be caused by sudden radio channel condition change in practice. In the simulation, we can introduce artificial obstacle, which results in sudden radio channel condition change. |

## SLS to evaluate HO performance

RAN2 agreed following principle for SLS:

**Agreements**

1. The System level performance (e.g. HO performance) evaluation is optional (i.e. companies can bring results if they chose).
   1. System level performance for measurement event prediction can be prioritized by companies if they chose to do it.

1. RAN2 will prioritize discussions on intermediate KPI discussion before discussing system level performance for the corresponding use case.

2: Discussion on what (type of information)/how generalization study can take place in Nov. meeting

3: The SLS simulation assumption discussion is covered in the post#127bis email discussion by assuming:

 The simulation assumptions agreed for measurement event prediction and RLF prediction is taken as baseline for SLS in principle

 The HO model in 36.839 is taken as baseline

 The HO performance will be HOF and number of HO only and definition in 36.839 is taken as baseline

 The baseline of HO performance is R15 legacy measurement and HO procedure

The following discussion starts with the baseline assumptions.

### Methodology

Here is Figure 5.2.1.3.1 in 36.839 to illustrate handover failure model in state 2:



Figure 2.3.1-1

In this Figure HO command is transmitted after receiving measurement report and the gap is “HO preparation time”. At least for FR2 temporal domain case A, it is expected that a predicted measurement event will be reported in advance. The reported measurement event can be used by the network to trigger preparation among source and target nodes so that the HO command can be transmitted in time.

For FR2 temporal domain case A, assuming the predication accuracy within PW is acceptable then the model can predict an A3 event with a timing advance roughly equal to PW. Considering the sampling period of FR2 is 80ms, which is higher than HO preparation time i.e. 50ms in 36.839, the HO preparation time can be saved completely. It means for case A, HO command is transmitted at the time where an A3 event is predicted. But Since ETD is allowed between real A3 event and predicted A3 event for “true event prediction”, it is also possible that a real A3 event is triggered before the predicted A3 event considering the OW is sliding step by step, where actual measurement is performed.



Figure 2.3.1-2 Example timeline for FR2 temporal domain case A based on direct prediction

In Figure 2.3.1-2, at current time instance t0, an A3 event at time instance t1 is predicted and reported. Based on this predicted event A3, network can finish handover preparation. However, when OW slides till time instance t2, a real A3 event can be triggered and reported and naturally network should start handover execution immediately even the candidate cell contained within the real A3 event report is different from previous reported predicted A3 event.

If direct prediction is adopted and no real A3 event is reported, then it is not clear when should network transmit handover command.



Figure 2.3.1-3 Example timeline for FR2 temporal domain case A based on indirect prediction

In Figure 2.3.1-3 the blue event occurrence window (assuming interpretation 1 in Figure 2.1.1-2) is an uncertain window of predicted A3 event. It seems the middle of the window i.e. t1 in Figure 2.3.1-3 is the compromised time for network to transmit HO command. If a real A3 event is triggered and reported early than t1 e.g. t2 in the Figure, then obviously network should transmit HO command at t2.

For both direct and indirect prediction, if the time to transmit HO command is early than the time point to finish HO preparation, then HO command is delayed until HO preparation is finished.

For FR1 temporal domain case B the situation is similar in the sense that measurement result can be also predicted to some extent for indirect prediction methodology. The difference compared to FR2 temporal domain case A is that t1-t0 could be smaller or even 0 ms. In addition, because at least some of the actual measurement results are skipped, there is no real A3 event is expected before predicted A3 event.

**The HO model for FR2 temporal domain case A is defined as following:**

**If a predicted A3 event at future t1 is reported at t0 and one potential real A3 event is triggered at t2, then HO command is transmitted at t3, where t3=t0+ max(HO prep time, min(t1-t0,t2-t0)). After that one fixed execution time is assumed. For direct prediction, t1 is the middle of the event occurrence window.**

**The HO model for FR1 temporal domain case B is defined as following:**

**If a predicted A3 event at t1 is reported at t0 (t0<=t1) then HO command is transmitted at t3, where t3=t0+max(HO prep time, t1-t0). After that one fixed execution time is assumed.**

**Question 22: Do you agree to the recommended HO model? If not, please provide you suggestion**

|  |  |  |
| --- | --- | --- |
| Company | Opinion: Yes or No | Comments |
| vivo | No | For indirect event prediction in Figure 2.3.1-2, t2 will always be after t0, so t3 should be equal to t0 + HO prep time. Only if when there is not a predicted A3 event reported at t0, t3 will be equal to t2 + HO prep time. |
| Xiaomi |  | In general, we want to clarify that NW can start the HO prepration after receiving the event prediction. After prediction, UE would still perform measurement to evaluate A3 event and eventually trigger MR if event is fulfilled. So, NW shall still wait for the real MR to send the HO. If NW sends the HO earlier than the real MR, handover too early would occur.  Based on above assumption, there are only two cases,  Case 1, HO preparation time > t2-t0, HO preparation is not finised at t2, so, t3 = t0 + preparation time  Case 2, HO preparation time < t2-t0, HO preparation has finised at t2, so, t3 = t2  In short, t3 = t0 + max (HO preparation time, t2-t0) |

### Metrics

RAN2 agreed that HO failure rate is metric of SLS. In order to calculate HO failure rate, we need know total HO number and the number of HO failures.

RAN2 also agreed that HO failure model in 36.839 is taken as baseline. Figure 2.3.1-1 is one of the criteria where T310 is running when HO command is transmitted. Here is the collection of the criteria in section 5.2.1.3 of 36.839:

**Definition 3:** A handover failure is counted if a RLF occurs in state 2, or a PDCCH failure is detected in state 2 or state 3.

For calculating the handover failures for the two states:

**- In state 2:** when the UE is attached to the source cell, a handover failure is counted if one of the following criteria is met:

1) Timer T310 has been triggered or is running when the HO\_CMD is received by the UE (indicating PDCCH failure) or

2) RLF is declared in the state 2

**- In state 3:** after the UE is attached to the target cell a handover failure is counted if the following criterion is met:

- target cell downlink filtered average (the filtering/averaging here is same as that used for starting T310) wideband CQI is less than the threshold Qout (-8 dB) at the end of the handover execution time (Table 5.1.4.1) in state 3.

…

**Definition 4**: The handover failure rate is defined as: Handover failure rate = (number of handover failures) / (Total number of handover attempts).

The total number of handover attempts is defined as: Total number of handover attempts = number of handover failures + number of successful handovers. The number of handover failures is in Definition 3.

Table 2.3.2-1 HO failure model

Note in order to judge whether T310 is running or expires when HO command is received, a hypothetical RLF procedure is necessary and SINR of the serving cell need be measured. In this case no RLF prediction is needed during SLS.

**Question 23: Do you agree to reuse HO failure model and corresponding metrics i.e. HO failure rate, total number of HO attempts from 36.839 as indicated in table 2.3.2-1?**

|  |  |  |
| --- | --- | --- |
| Company | Opinion: Yes or No | Comments |
| vivo | Yes for HO failure rate,  No for number of HO attempts | For the number of HO attempts, the duration of the simulation may be different. Therefore, suggest refining it as ‘Total number of handover attempts per UE per second’. |
| Xiaomi | Yes for HOF rate  No for HO number | Agree with vivo. The HO number should be calculated per second. |

### Simulation assumptions

By answering the questions in section 2.1 and 2.2, we can get simulation assumptions for measurement event prediction (both FR1 and FR2). In principle those parameters can be reused for SLS.

**Question 24: Do you agree that parameters concluded in section 2.1.3 for measurement event prediction can be reused for SLS? If not, please pointed out which parameters need updated**

|  |  |  |
| --- | --- | --- |
| Company | Opinion: Yes or No | Comments |
| vivo |  | The same comments in section 2.1.3. |
| Xiaomi | Yes | Only the agreeable part |

For RLF only partial parameters (i.e. all but Max ETD, uncertain window length and probability threshold) in table 2.2.2-1 is needed because no RLF prediction is needed in SLS. Another issue is whether the same set of parameters can be also used for FR1 considering SLS can cover FR1 temporal domain prediction also.

**Question 25: Do you agree that parameters in table 2.2.2-1 i.e. all but last 3 parameters can be reused for both FR2 temporal domain case A and FR1 temporal domain case B in SLS? If not, please elaborate parameters for FR1**

|  |  |  |
| --- | --- | --- |
| Company | Opinion: Yes or No | Comments |
| vivo | Yes |  |
| Xiaomi | Yes |  |

For RLF, the interference modelling in section 2.2.2 can be also reused for SLS.

**Question 26: Do you agree interference model in section 2.2.2 can be reused for SLS? If not, please pointed out which parameters need updated**

|  |  |  |
| --- | --- | --- |
| Company | Opinion: Yes or No | Comments |
| vivo | Yes |  |
| Xiaomi | Yes |  |

As discussed in section 2.3.1 HO preparation time and execution time is need to model HO procedure. the value in 36.839 is 50ms and 40ms respectively. The question is whether these two parameters can be reused and need be differentiated between FR1 and FR2.

**Question 27: How do you think of HO preparation time and HO execution time for FR1 and FR2 respectively?**

|  |  |  |
| --- | --- | --- |
| Company | Opinion: Yes or No | Comments |
| vivo | Yes | Assumptions in TR 36.839 can be reused |
| Xiaomi |  | Can be the same for FR1 and FR2 |

## Inter-frequency correlation coefficient

In RAN2#127bis, it is agreed that companies should report with their simulation the correlation coefficient. The correlation coefficient is intended to show how much the L3 RSRP of the two cells in different frequencies are correlated in simulation. It can help reflect whether good prediction accuracy results from AI or from a stronger correlation.

It should be noted that correlation coefficient is not a parameter defined by TS 38.901 but rather a calculated value. There are several ways to get the coefficient such as Spearman correlation coefficient and Kendall correlation coefficient. In [14], PCC (Pearson Correlation Coefficient), which indicates the correlation between two vectors as a scalar value from -1 to 1, is used for the calculation. For inter-frequency RRM measurement prediction, the two vectors can be “series of RSRP in multiple time instances of the serving cell” and “series of RSRP in the same multiple time instances of the co-sector cell”. The series of L3 RSRP values of the two cells can be collected within the same time window, but it can be up to companies how to set the time window in their simulation (e.g., whole simulation time or any time window during simulation time). There are existing functions in MATLAB (correcoef(x,y)) and Python (pearsonr(x,y)) for PCC calculation. Given its simplicity, PCC seems to be a good choice.

**Question 28: Do you agree to use Pearson correlation coefficient for correlation coefficient calculation?**

|  |  |  |
| --- | --- | --- |
| Company | Opinion: Yes or No | Comments |
| vivo | Yes |  |

# Conclusion

# Reference

[1] R2-2407978 Discussion on measurement event prediction OPPO

[2] R2-2406705 Discussion on RLF prediction assumptions Xiaomi

[3] R2-2409011 Text proposal on TR 38.744 OPPO draft TR Rel-19 38.744 0.0.4 FS\_NR\_AIML\_Mob

[4] R2-2408737 Discussion on simulations for measurement event prediction Huawei, HiSilicon discussion Rel-19 FS\_NR\_AIML\_Mob

[5] R2-2407978 Discussion on measurement event prediction OPPO discussion Rel-19 FS\_NR\_AIML\_Mob

[6] R2-2409207 Evaluation on RRM measurement prediction ZTE Corporation discussion Rel-19 FS\_NR\_AIML\_Mob

[7] R2-2406976 Discussion on Simulation Assumption and Methodology for RLF prediction CMCC

[8] R2-2407492 Discussion on simulation assumptions and evaluation methodology for RLF prediction Samsung

[9] R2-2407093 AI-ML based RLF predictions Ericsson

[10] R2-2407211 R19 NR AIML A8341\_RLF\_prediction\_simulation\_assumptions Interdigital

[11] R2-2407481 RLF Prediction Aspects Nokia

[12] R2-2407389 Simulation assumption for RLF prediction KDDI

[13] R2-2406312 Discussion on RLF use case OPPO

[14] R2-2408529 Discussion on the simulation results for RRM measurement Samsung

# Annex: RAN2 agreements in #127 meeting

Table 5-1

**Agreements**

- Both direct and indirect are allowed. Companies should indicate what they used and what inputs they are using

- Output for indirect: predicted SINR. Based on predicted SINR the time instance the RLF occurs can be determined without further AI/ML models.

- Output for direct: probability of RLF within an window

- Companies should report the prediction time window they have used in the simulations

**Agreements on KPI**

For indirect

- SINR difference, missed RLF detection, false RLF detection, F1 score, time difference of true time RLF and predicted RLF, true RLF prediction.

Direct

- missed RLF detection, false RLF detection, F1 score.

- Whether the RLF will happen based on the following methodology. Above a probability threshold we assume that RLF will happen. This is then compared with true RLF.

- For the time being we don’t need HO procedure simulation in RLF simulation

- FFS full buffer and assumption that all the cells are fully loaded. We will not simulate traffic.

- Simulation results are not expected before February

# Annex: RAN2 agreements in #127bis meeting

**Agreements**

* Measurement event prediction simulations will at least focus on intra-frequency FR2, case A, and second study goal (i.e. HO KPI improvement). FFS what is KPI.
* Companies can bring simulation results for intra-frequency measurement reduction for FR1 and report what they are doing. Focus on temporal case B.
* Companies will prioritize simulations on indirect method. Companies can bring simulations on direct method and should report what method is being used.
* Measurement event prediction results are expected in RAN2#129

Agreements on simulation assumptions

* *The Simulation assumption of RRM measurement prediction can be reused unless otherwise specified.*
* Companies can pick and report what they are using for filtering options (similar to RRM prediction)
* Companies will focus on sub-case 2 for measurement event prediction. Companies can simulate other sub-cases if they wish and report what they are using.
* Leave the simulation parameter discussion for email discussion. Pick only one value for A3.
* It is up to company implementation how to model UE behavior after A3 event is trigger. Focus on intermediate KPIs for this exercise. System level KPI is FFFs
* For measurement event prediction, traffic is not simulated.

**Agreements on inputs/outputs and KPIs**

1. For indirect measurement event prediction, the intermediate output (i.e., the output of RRM prediction model) is RSRP of serving/neighboring cells. The final output is the expected occurrence time of a certain measurement event (ex. event A3).
2. For direct measurement event prediction, the model output is the probability of event occurrence within a time window.
3. A3 event prediction should follow legacy rules (i.e. the “predicted” conditions have to persist for the duration of TTT).
4. As baseline, we will use RLF event prediction KPI:

for indirect: F1 score. the following can be reported: RSRP difference, missed event detection, false event detection. FFS how to define F1 score.

time difference of true time event reporting triggered and predicted time event reporting triggered, true event prediction.

for direct: F1 score. The following can be reported: missed event detection, false event detection,

Continue discussion over email discussion to see if there is a difference.