3GPP TSG RAN Meeting #102
Edinburgh, Scotland, December 11 - 15, 2023
Agenda: 9.1.2.4
Document for: Discussion
RP-233576

Views on AI/ML-based mobility enhancement

China Telecom, vivo

December 2023

Background

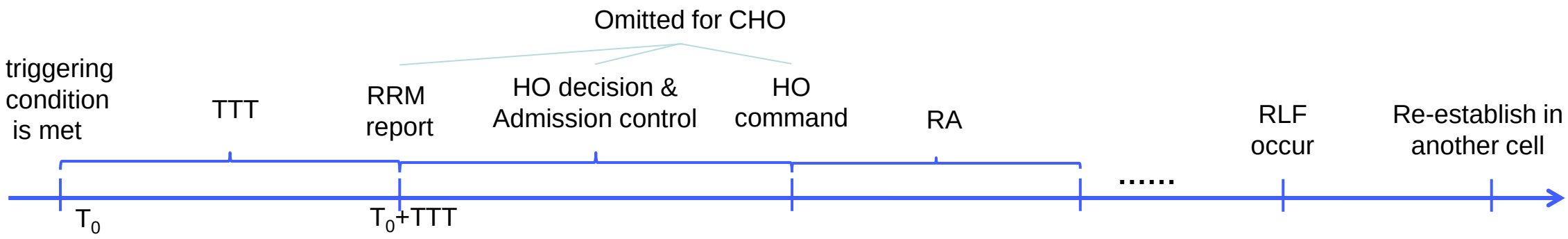
- **In the summary for RAN Rel-19 Package in RP-232745, the potential objectives for AI for for Air Interface (Mobility) SI were captured as below:**

  - References: RWS-230488, RP-231540, RP-232622
  - Potential objectives: possible down-scoping in RAN#102
    - Type of mobility
      - L3-based mobility and L1/L2-triggered mobility (LTM) are both considered
    - HO optimization in Network side ~~[/UE side]~~, including
      - Candidate/target cell prediction in L3-based mobility, or, candidate/target beam(s) and cell(s) prediction in LTM
    - RRM measurement and event prediction, including
      - Beam-level measurement prediction
      - Cell-level measurement prediction, e.g., using intra-frequency measurement results to forecast the RRM measurement of inter-frequency/inter-RAT cells
      - HO failure/RLF prediction
      - Measurement events prediction
    - LCM framework and others
      - The conclusions in Rel-18 AI/ML study should be used as baseline
      - Other impacts are further studied
    - ~~[UE RRM enhancement, RAN4]~~
    - Note 1: no intention to change the existing framework for the mobility under connected mode
    - Note 2: In the SID, target performance metrics and impacts should be clarified
    - Note 3: Avoid overlap work with RAN3

Background

- In Rel-18, the *NR_AIML_NGRAN-Core* WI specifies the AI/ML based mobility optimization:
  - The Model Inference functionality resides within the RAN node only supporting cell-level mobility prediction.
  - Some location-related information of UE (e.g., coordinates) was required as the input of AI/ML model, which may introduce UE privacy concerns. As a result, UE may choose not to provide these information.
  - Frequent exchange of the input between network and UE will introduce massive signaling overhead and latency.
- In Rel-18, the *NR_AIML_Air* SI would introduce the model life cycle management to enable the use cases at the air interface:
  - The framework can be reused to facilitate more use cases at the air interface.
  - The Model Inference functionality can reside on the UE side.
  - Local model inference at UE side may utilize more detailed location information and can reduce signaling overhead of input exchange.
- **Observation 1: For mobility optimization, if the Model Inference functionality can be deployed on the UE side:**
  - **More detailed local information from UE can be utilized as the input to improve prediction accuracy without privacy concerns.**
  - **Local model inference can reduce signalling overhead and inference latency.**

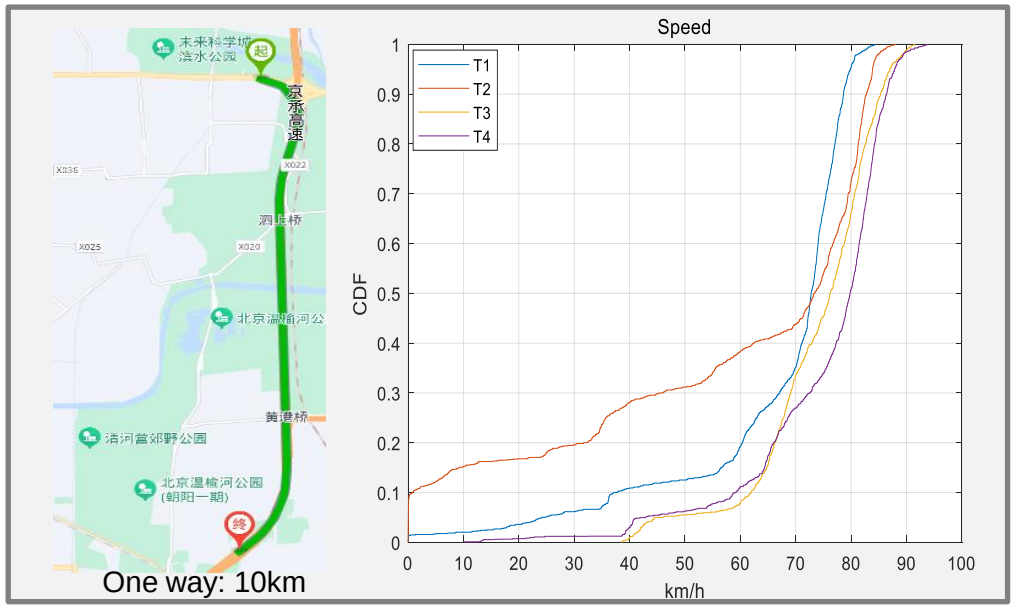
Issues and Motivations

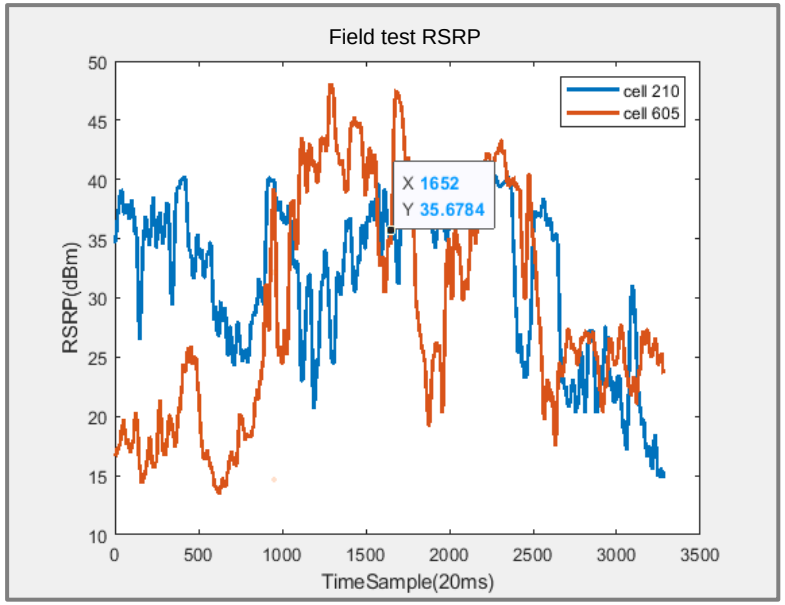


- For the legacy HO, the triggering condition for RRM reporting shall be met at T_0 and shall last for TTT (Time To Trigger) duration, which may lead to:
  - HO at a non-optimal time, poor user experience at source cell.
  - Failed to receive the HO command or failed to RA to the target cell, i.e., too-late HO.
- For CHO, the UE can RA to the target cell without receiving the HO command if the triggering condition is met during TTT. However, there is still a risk of RLF due to low SINR at the source cell during TTT and the UE cannot perform HO at a optimal time.
- The legacy solution to the above issue by reducing TTT duration may result in other unintended events, e.g., too-early HO, ping-pong HO, especially for the high-speed UEs.
- If an RLF occurs shortly after a successful HO, the UE may attempt to re-establish the radio link connection in a cell other than the source cell and the target cell, which is identified as HO to wrong cell.

Issues during field test - Highway

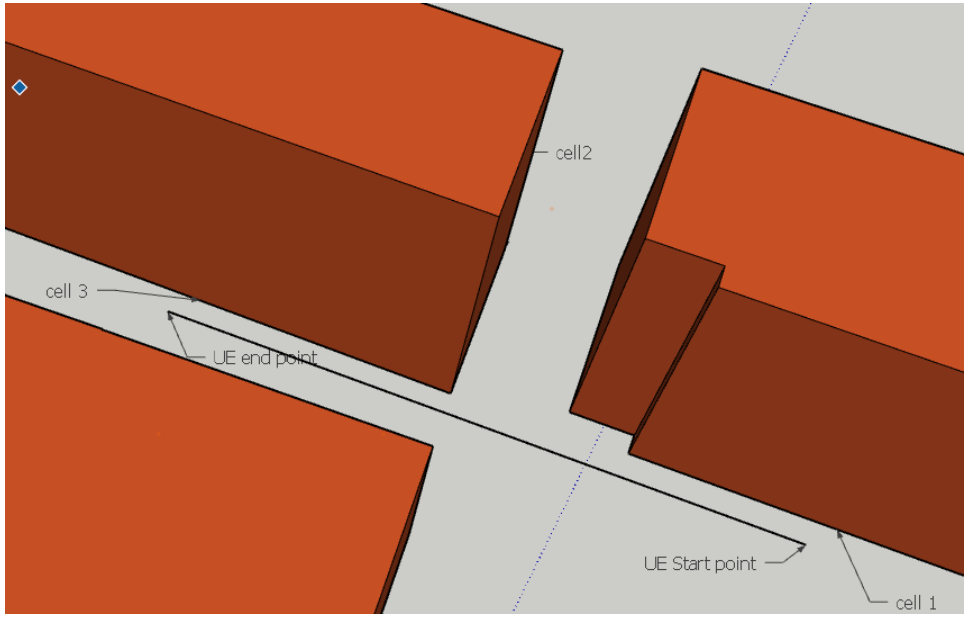
Carrier Frequency: FR1, 3.5GHz，UE speed



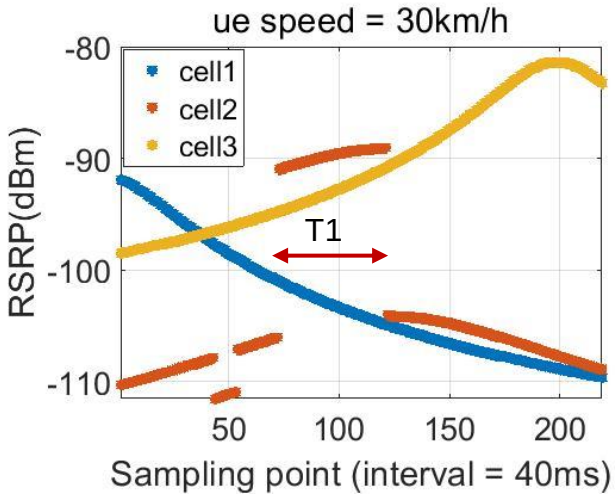


- During UE high-speed movement, the RSRP of the neighbor may becomes better than the serving cell in a short period of time.
- UE will handover to the neighbor cell and handover back quickly to the last serving cell, i.e., ping-pong handover occurs.

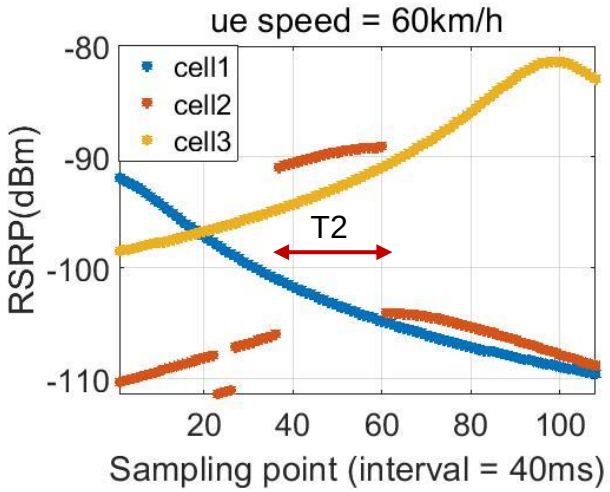
Issues found by simulation - Crossover



Simulation scenario and UE trajectory



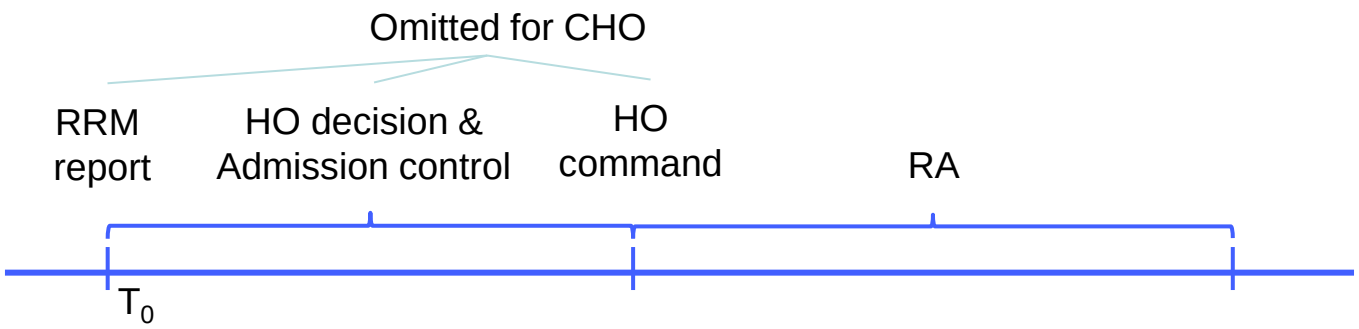
UE1 speed: 30km/h
T1 duration: 1960ms



UE2 speed: 60km/h
T2 duration: 960ms

- When the UE passes through the crossover, the RSRP of cell2 will change dramatically and UE handover to cell 2.
- For UE1 at a low speed, it will be served by cell 2 for a longer time (over 1 second).
- For UE2 at a high speed, it will only be served by cell 2 for less than 1 second.
- Both UEs may experience RLF when leaving the crossover and will reestablishment RRC connection on cell 3.

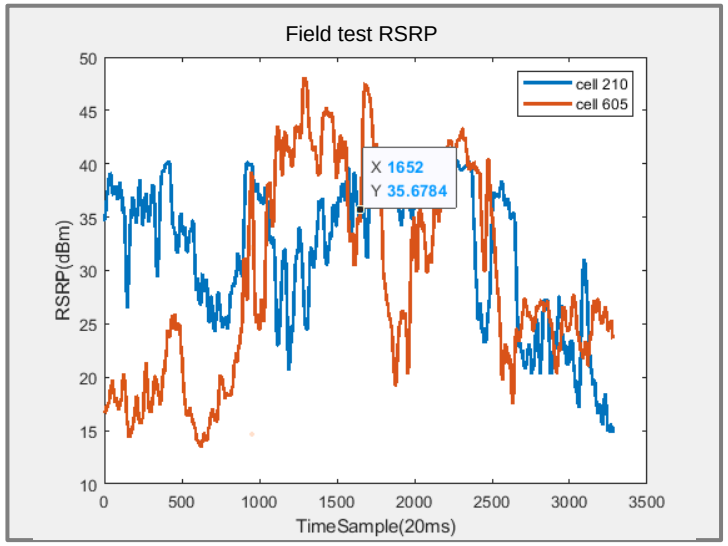
Potential solution



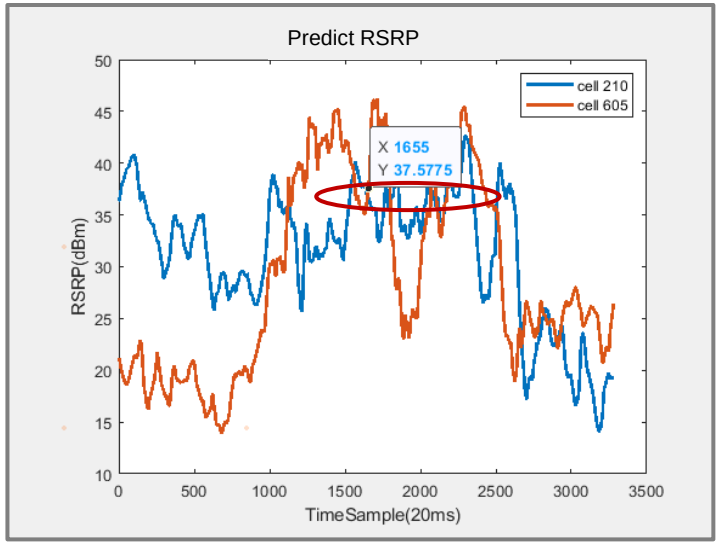
- UE may perform RRM prediction, e.g. RSRP prediction.
- For the legacy HO, with RSRP prediction, the UE can send the RRM report once the triggering condition is met at T_0 (i.e., no need to wait for TTT duration).
  - RSRP prediction within the T_0 +TTT period shall meet the triggering condition,
  - The UE may send the RSRP prediction of neighbor cells during RA in the RRM report for HO decision,
  - Higher success rate for receiving HO command when the RRM report was sent at an optimal time.
- For CHO, with RSRP prediction, the UE can RA to the target cell once the triggering condition is met at T_0.
  - RSRP prediction within the T_0 + TTT period shall meet the triggering condition,
  - Lower risk of RLF at the source cell when the HO is performed at a optimal time.
- During the HO decision and target cell selection, the RSRP prediction can be used to reduce the unintended events, e.g., ping-pong HO, too early HO, HO to wrong cell.

Feasibility analysis with field test - Highway

- **Field test RSRP**



- **Predict RSRP**



- **UE speed: 0~100km/h**
- **AI/ML Model**
  - Fully Connected Neural Network
- **Input**
  - History RSRP of serving & neighbor cell (interval = 20ms)
  - History UE location and speed (interval = 1s)
  - Observation window = 2s
- **Output**
  - Predict RSRP of serving cell
  - Predict RSRP of neighbor cell
  - Prediction timing = 2s
- **Performance [2][3][4][5]**
  - RMSE (root mean square error) < 3~4 dB
  - the ping-pong HO rate and short time of stay were reduced by more than 94%
  - Field test results show that the predicted RSRP is basically consistent with the filed test RSRP

Feasibility analysis with simulation – Crossover

- **Accuracy of RRM measurement prediction [1][2][3][4]**

	Cell 1	Cell 2	Cell 3
Prediction 1	RMSE = 0.0044dB	RMSE = 1.08dB	RMSE = 0.26dB
Prediction 2	RMSE = 0.0062dB	RMSE = 1.11dB	RMSE = 0.26dB
Prediction 3	RMSE = 0.0844dB	RMSE = 1.23dB	RMSE = 0.28dB

Carrier Frequency: FR2, 30GHz

Prediction 1: RSRP of every 80ms in 320ms after T_0

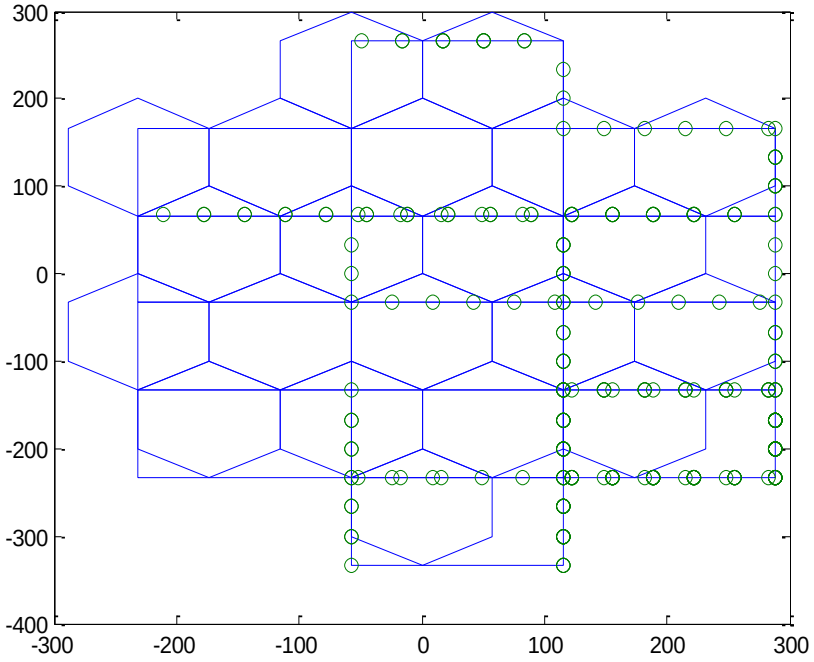
Prediction 2: RSRP of 1s after T_0

Prediction 3: RSRP of 2s after T_0

Simulation results show that the RRM measurement predictions on different time scale have very low RMSE.

Performance analysis with simulation

Simulation scenario



Simulation scenario and UE trajectory

Simulation assumption

Attributes	Values or assumptions
Carrier Frequency	FR1: 4GHz; FR2: 30GHz
TRP Number	7 sites, 3 sector per site
Channel Model	3D-Uma in TR 38.901, support Spatial consistency ISD = 200m
UE speed	120km/h
Mobility management	Event: A3; Hysteresis: 2dB; Offset: 1dB; TimeToTrigger: 320ms, 40ms Handover preparation time: 50ms; Handover execution time: 40ms
RLM	L1 measurement period: 20ms Qin sliding window length: 100ms Qout sliding window length: 200ms Qin threshold: -6dB; Qout threshold: -8dB N310: 1; N311: 1; T310: 1s
Handover model and corresponding metrics	As defined in TR 36.839 Short time of stay: served by the target cell for less than 1s after HO

Performance analysis with simulation

Accuracy of RRM prediction [1][2][3][4]

	Prediction 1	Prediction 2
FR1	RMSE = 0.38 dB	RMSE = 1.6 dB
FR2	RMSE = 1.3 dB	RMSE = 3.3 dB

Training dataset: Same large scale channel parameters for different drops

Prediction 1: RSRP of every 80ms in 320ms after T_0

Prediction 2: RSRP of 1s after T_0

Simulation results shows that the RRM measurement predictions on FR1 and FR2 have reasonable RMSE.

Usage of RRM prediction

- Legacy HO:
  - UE can decide whether to trigger the RRM reporting based on the predicted RSRP of every 80ms in 320ms after T_0,
  - Source cell can determine the target cell based on the predicted RSRP of 1s after T_0 to avoid too early HO or HO to wrong cell,
  - Source cell can forward the predicted RSRP to target cell for admission control.
- CHO:
  - UE can decide to trigger the target cell selection based on the predicted RSRP of every 80ms in 320ms after T_0,
  - UE can choose the target cell based on the predicted RSRP of 1s after T_0 to avoid too early HO or HO to wrong cell.

Performance analysis with simulation

- **RRM prediction based HO [1][2][3][4]**

		Legacy HO, TTT = 320	Legacy HO, TTT = 40	AI/ML based HO	CHO, TTT = 320	CHO, TTT = 40	AI/ML based CHO
FR1	HOF rate	9.16%	2.2%	1.95%	0.28%	0.15%	0.32%
	Ping-pong HO rate	1.1%	3.6%	0.37%	1.0%	3.7%	0.37%
	Short Time of Stay (1s) rate	13.4%	18.9%	5.7%	13.6%	18.8%	5.67%
FR2	HOF rate	7.4%	2.5%	2.0%	0.42%	0.43%	0.44%
	Ping-pong HO rate	5.2%	10.3%	2.7%	5.2%	10.3%	2.7%
	Short Time of Stay (1s) rate	24.1%	36.7%	10.4%	24.4%	36.5%	10.8%

- **Observation: With RRM prediction, the unintended events rate during HO and CHO can be significantly reduced, including HOF rate, ping-pong HO rate and short time of stay rate.**

Discussion on type of mobility and model sidedness

- **For the RRC state of mobility, in addition to HO in RRC_CONNECTED, Cell Reselection in RRC_INACTIVE/IDLE can also be considered.**
  - If the UE reselects to an unsuitable cell and is triggered to set up a connection with the camping cell, it may need to be HO or redirected to another suitable cell, which would impact user experience.

- **Conditional HO is also included in L3-based mobility.**
  - The current potential scope of L3-based mobility is not clear, i.e., whether both legacy HO in Rel-15 and Conditional HO are included. From our understanding, conditional HO-specific solution is foreseen during potential impact analysis.

- **For model sidedness of inference, UE-side model should be the baseline, NW-side model should avoid overlap with AI for NG-RAN in RAN3, and two-sided model can also be considered.**
  - During the initial evaluation, it is seen that data from NW and UE are essential to ensure AI/ML performance. The data from NW includes the location and beam configuration of NG-RAN node. If operators or NW vendors have concerns about the exposure of information to UE for UE-side model, two-sided model can also be considered.

Suggested Potential Objective

- Study solutions for AI/ML-based mobility optimization considering the following:
  - Type of mobility: L3-based (legacy HO and CHO) and L1/L2-based (LTM) mobility in RRC_CONNECTED, Cell Reselection in RRC_INACTIVE/IDLE .
  - Type of model:
    - UE-side model (baseline), NW-side model and two sided model
    - NW-side model should avoid overlap with AI for NG-RAN in RAN3;
  - Study and finalize the sub-use cases based on the performance evaluation of AI/ML based algorithms [RAN2, RAN1, RAN4]
    - Consider the following potential sub-use cases (To be down-selected at RAN#104):
      - HO optimization based on candidate/target cell prediction in L3-based mobility, or candidate/target beam(s) and cell(s) prediction in LTM, and considering the associated HO timing prediction;
      - Cell-level/beam-level RRM measurement (i.e., RSRP/RSRQ/SINR) prediction of serving/neighbor cells in the temporal/spatial/frequency domain;
      - Unintended events prediction, e.g., HOF, RLF, too-early HO, too-late HO, HO to wrong cell;
      - Measurement events prediction, e.g., event A3.
    - Study the input/output and corresponding mechanism of each sub-use case.
    - Evaluate and compare different sub-use cases with determined intermediate KPIs and system-level performance with methodology based on statistical models (from [TR 38.901]) for system-level simulations.
  - Assess potential specification impact for the agreed sub-use cases, including: [RAN2, RAN1, RAN3]
    - Identify the applicability and essential enhancement of AI/ML LCM framework in Rel-18 AI/ML;
    - Study the potential procedure and signaling exchange to facilitate the identified mechanism of each sub-use case.
- Note 1: no intention to change the existing framework for mobility under connected mode.
- Note 2: Focus on Standalone (SA) mode during SI phase, i.e., DC is not in the SI scope.

Conclusion

In this contribution, we have following observations and proposals:

- Observation 1: For mobility optimization, if the Model Inference functionality can be deployed on the UE side:
  - » More detailed local information from UE can be utilized as the input to improve prediction accuracy without privacy concerns.
  - » Local model inference can reduce signalling overhead and inference latency.
- Observation 2: With RRM prediction, the unintended events rate during HO and CHO can be significantly reduced, including HOF rate, ping-pong HO rate and short time of stay rate.
- Agree on a RAN2 leading AI/ML-based mobility enhancement study item in Rel-19.
  - » For the RRC state of mobility, in addition to HO in RRC_CONNECTED, Cell Reselection in RRC_INACTIVE/IDLE can also be considered.
  - » Conditional HO is also included in L3-based mobility.
  - » For model sidedness of inference, UE-side model should be the baseline, NW-side model should avoid overlap with AI for NG-RAN in RAN3, and two-sided model can also be considered.

References

In this contribution, we have following references for corresponding evaluation results:

[1] RP-230377, Study on AI based mobility enhancement in Rel-19, vivo

[2] RWS-230063, Views on Rel-19 AI/ML for air interface, vivo

[3] RWS-230240, Motivations on AI/ML-based mobility enhancement, China Telecom, vivo

[4] RP-231789, Rel-19 items for AI/ML over Air interface (RAN1-led WI and RAN2-led SI), vivo

[5] RP-232278, Views on NR Rel-19 AIML for air interface, China Telecom

Thanks!