**3GPP TSG-WG SA2 Meeting #154 *S2-220xxxx***

**Toulouse, France, November 14 – 18, 2022 (revision of S2-220xxxx)**

**Source: Huawei, HiSilicon**

**Title: KI#1: Update of Conclusions**

**Document for: Approval**

**Agenda Item: 9.23**

**Work Item / Release: FS\_eNA\_Ph3 / Rel-18**

*Abstract: This contribution proposes the update of the conclusions of KI#1.*

# 1. Introduction/Discussion

There is an EN in the conclusion of KI#1 as following:

*Editor's Note: The analytics consumer NF making some decision may change the trend indicated by the prediction output. The analytics consumer NF may provide a unified feedback related to the effect of an analytics on the changes in network status after the consumption of analytics. How to define such unified feedback and based on which logic is FFS.*

As the examples listed in Table 1, the analytics consumer, e.g. PCF or SMF, will take actions after getting prediction analytics, especially when there will be congestion or overload as indicated by the analytics result. In detail, when there is a prediction of the congestion, the network may take action to reduce the authorized bitrate. Then the potential network status will be “not congested”. So it is not aligned with the statistics. But it actually a good prediction as if there is no actions in 5GC, it will be congested. That’s why NF feedback is required because only the consumer know about the NF action and whether the result with taking the input of NWDAF into account is better or worse compared to the one without input of NWDAF. Other examples are similar.

Table 1: Possibility of trend change after actions determined based on prediction analytics

|  |  |  |  |
| --- | --- | --- | --- |
| **Analytics IDs** | **Analytics Consumer NF** | **Possible actions after getting prediction analytics** | **Whether the ground truth is vulnerable to deviate from prediction after actions** |
| User Data Congestion | PCF | Based on the "User Data Congestion" predictions including the list of applications contributing the most to the traffic the PCF may perform SM Policy Association modifications to update policies in the SMF for the PDU sessions handling traffic from those applications. | Yes.  The top applications are easy to change after suppressing by the PCF. |
| NF load information | OAM | Based on the NF load predication, OAM may allocate more NFV resources to the 5GC NF to prevent overload. | Yes.  The NF resource usage in the predication is easy to change after allocating more NFV resources. |
| (Slice) Load level information | NSSF/AMF | Slice load analytics information may be used for the determination of the Network Slice instance(s) and the list of S-NSSAI(s) in the Allowed NSSAI(s) to serve the UE. | Yes.  The Number of UE Registrations in the predication are easy to change after NSSF/AMF stopping to select the network slice instance. |
| Data Dispersion | PCF | AM Policy Association modification to update UE-AMBR, RFSP index and/or service area restriction, for those UEs reported as heavy users.  SM Policy Association modification to update the policies in the SMF for those UEs reported as heavy users. | Yes.  The top-heavy users in the predication are easy to change after decreasing the authorized UE-AMBR or session-AMBR. |
| DN Performance | SMF | SMF (re)selects UP paths (including DNAI(s)) for PDU Sessions.  SMF configures traffic steering at UPF. | Yes.  The Average traffic rate predicted of a DNAI is easy to change after traffic steering. |
| Network Performance | PCF | The PCF may use the "Network Performance" predication as input to calculate the background data transfer policies that are negotiated with the ASP. | Yes.  The gNB resource usage predicted is easy to change after background data transfer. |
| WLAN performance | PCF | Based on the WLAN performance statistics or predictions, the PCF may update WLANSP. | Yes.  The number of UEs predicated is easy to change after WLANSP update. |
| Redundant Transmission Experience | SMF | SMF may make use of Redundant Transmission Experience analytics, when SMF takes a decision whether to perform redundant transmission, or stop redundant transmission if it had been activated | Yes.  The ratio of UE efficiently use the PDU session with redundant transmission in the prediction is easy to change after action. |
| Service Experience | PCF/UPF | The PCF, based on service experience analytics per UP path, determines for each DNAI, a traffic steering policy ID.  SMF may use Service Experience analytics per UP path to select UPF. | Yes.  The service experience predicated for an Application Server Instance Address is easy to change after switch traffics to another path. |
| UE Mobility | AMF | AMF may use UE mobility analytics for RA decision and paging. | Not easy.  The UE trajectory is an inherent feature of the UE. The UE location predicated is not easy to change by the action of the network. |
| UE Communication | SMF | SMF may use UE related analytics (UE mobility, UE communication, and expected UE behavioural parameters) for UPF selection. | Not easy.  The UE communication action is an inherent feature of the UE. The UE communication predicated is not easy to change by the action of the network. |
| QoS Sustainability | AF | Application Server may request QoS Sustainability prediction for the purposes of adjustment of the application. | Not easy.  Only one application changing the communication mode is not easy to change the QoS sustainability of the network, especially when the application is not the top-heavy application. |
| Abnormal behaviour | AMF | Currently there is no use case in the real network using the predication of abnormal behaviour. | No use case to judge. |
| Session Management Congestion Control Experience | SMF | Currently there is no definition of the predication output. | No use case to judge. |

**Observation 1: ML Model accuracy improvement cannot be achieved by comparing prediction and its corresponding ground truth data for more than half Analytics IDs in TS 23.288, especially when there will be congestion or overload as indicated by the analytics output.**

In order to avoid foreseeable congestion or overload as described in the above example, the analytics consumer takes actions based on some local policies, which include performance objectives and local service logic. If the prediction is correct, the actions taken based on the prediction will lead to expected performance objectives according to the local service logic. The NF knows the gain/output of local service logic as there must be an internal algorithm on . The analytics consumer tells whether the prediction is useful or not based on the difference between the KPIs related to the performance objectives without/before taking the analytic into account and the one with taking the analytic into account.

There are some detailed examples are listed in the following Table 2.

Table 2: Examples of judging the correctness of analytics prediction

|  |  |  |
| --- | --- | --- |
| **Analytics IDs** | **Analytics Consumer NF** | **How can consumer NF make the decision on the feedback** |
| User Data Congestion | PCF | The PCF gets the throughputs before action and after suppressing the top-heavy applications. The PCF make sure the throughputs are controlled based on e.g. Resource allocation outcome, QNC, etc, if the Top-Heavy applications listed in the "User Data Congestion" prediction are suppressed. |
| NF load information | OAM | The OAM monitors the NF resource usage before action and after action. The OAM tells whether the NF resource usage is changed to a proper range after allocating certain NFV resources based on the predication. |
| (Slice) Load level information | NSSF | The NSSF gets the load of SMF and UPF in the network slice instances before action and after adjusting the distribution proportion among network slice instances. The NSSF tells whether loads of the NFs in the slice instance is controlled in a proper range (i.e. no overload or inefficiency) based on the predication. |
| Data Dispersion | PCF | The PCF gets the network performance of the AOI before action and after decreasing the authorized UE-AMBR or session-AMBR. The PCF judges whether the network performance of a certain area is controlled after suppressing top-heavy users based on the predication. |
| DN Performance | SMF | The PCF gets data volume dispersion statistics bound by location before action and after traffic steering and judge whether load distribution proportion among DNAIs are controlled based on the predication. |
| Network Performance | PCF | The PCF monitors the gNB resource usage before planed background data transfer and tells whether the gNB resource usage is controlled based on Resource allocation outcome during background data transfer. |
| WLAN performance | PCF | The PCF gets the average throughput per UE before action and after WLANSP update. The PCF tells whether the performance of the WLAN is controlled based on the change of the average throughput per UE. |
| Redundant Transmission Experience | SMF | The PCF monitors the service experience and resource usage before action and after action. The PCF tells whether the changes of both the service experience and the resource usage are in a proper range. |
| Service Experience | PCF/UPF | The PCF monitors the service experience before action and after load balance adjustment, and tells whether both the load balance between Application Server Instance Addresses and the average experience are controlled in a proper range. |
| UE Mobility | AMF | No need to provide feedback |
| UE Communication | SMF | No need to provide feedback |
| QoS Sustainability | AF | No need to provide feedback |
| Abnormal behaviour | AMF | No need to provide feedback |
| Session Management Congestion Control Experience | SMF | No need to provide feedback |

Besides, The analytics consumer may get statistics of another Analytics ID, of which the output includes the KPIs related to the goal of the action, and judge whether there are positive effects because of the actions. For example, the PCF modifies SM Policy Association to decrease the traffic of Top-Heavy applications listed in the "User Data Congestion" prediction, and consume the "Data Volume Dispersion" statistics across all the applications in the AOI to make sure the throughputs are controlled in a proper range. If the Top-Heavy applications listed in the "User Data Congestion" prediction are correct, the throughputs should decrease as expected after actions. The PCF can tell whether the actions taken based on the prediction analytics have positive effect on the goal of the action.

**Observation 2: The analytics consumer can tell whether the prediction is useful or not based on the difference between the KPIs related to the performance without/before taking the analytic into account and the one with taking the analytic into account.**

In Nnwdaf\_AnalyticsInfo\_Request Response or Nnwdaf\_AnalyticsSubscription\_Subscribe Response the AnLF may instruct the analytics consumer to provide feedback about the prediction result of an Analytics ID, which is an implicit subscription to the notification of accuracy feedback. If the analytics consumer makes sure that the consumption of the prediction analytics had positive effect on network performance, the analytics consumer provides to the AnLF a "hit" indication as the feedback. Otherwise, the analytics consumer provides to the AnLF a "not hit" indication. For example, the PCF may perform SM Policy Association modifications to decrease the traffic of Top-Heavy applications listed in the "User Data Congestion" prediction. If the Top-Heavy applications listed in the "User Data Congestion" prediction are correct, the congestion will be successfully avoided. The PCF provides to the AnLF a "hit" indication in the accuracy feedback notification message.

If the AnLF has registered in MTLF with its accuracy monitoring for the model, the AnLF uses the feedback in the accuracy check for the respective Analytics ID. Otherwise, the unified feedback from the service consumer to the AnLF needs to be forwarded to the MTLF, and the MTLF uses the feedback in the accuracy check for the respective Analytics ID. If the feedback from most of the analytics consumer indicates "not hit", the MTLF may need to reselect a new ML model or retrain the existing ML model provided to the AnLF.

**Observation 3: The analytics consumer can be triggered to provide the unified feedback of an Analytics ID by the implicit subscription in Nnwdaf\_AnalyticsInfo\_Request Response or Nnwdaf\_AnalyticsSubscription\_Subscribe Response from AnLF. The AnLF may forward unified feedback of an Analytics ID to the MTLF. The AnLF or the MTLF determines ML model degradation related to the Analytics ID if the unified feedbacks of the Analytics ID from most of the analytics consumers indicate "not hit".**

According to the above observations, the following proposal is provided.

**Proposal: AnLF or MTLF (via AnLF) should get unified feedbacks of an Analytics ID from the prediction analytics consumer, which is "hit" or "not hit" indication according to the difference between the KPIs related to the performance without/before taking the analytic into account and the one with taking the analytic into account**. **The analytics consumer is triggered to provide the unified feedback of an Analytics ID in accuracy feedback notification message by the implicit subscription in Nnwdaf\_AnalyticsInfo\_Request Response or Nnwdaf\_AnalyticsSubscription\_Subscribe Response.**

So, the above EN in the conclusion of KI#1 can be resolved.

# 2. Text Proposal

It is proposed to capture the following changes vs. TR 23.700-81.

\* \* \* \* First change \* \* \* \*

## 8.1 Key Issue #1: How to improve correctness of NWDAF

For KI#1, it proposes the following principles:

**General aspects:**

- Analytics consumers and AnLF may indicate a "Use case context" when subscribing to or requesting analytics or ML model(s), respectively. The values of this parameter will not be standardized. The actions of the NWDAF based on the use case context are out of scope of 3GPP/implementation specific.

- NWDAF has the accuracy checking capability of analytics IDs and/or ML models, where NWDAF can store for a period of time the necessary information to determine the analytics IDs and/or ML model accuracy and provide the accuracy information to consumers when requested or use it for its internal processes.

- An NWDAF containing AnLF NWDAF with accuracy checking capability is able to provide or notify the accuracy information of Analytics IDs to the consumers of such service.

- An NWDAF containing MTLF NWDAF with accuracy checking capability is able to provide or notify the ML model accuracy degradation to the consumers of such service.

Editor's Note: It is FFS if consumers of AnLF and MTLF services can generate and request storage of performance information in such a way that other service consumers can retrieve it.

**Input of** accuracy **check:**

- ML Model accuracy improvement can be achieved by comparing prediction using the current trained ML model and its corresponding ground truth data i.e. the corresponding true observed events.

- The MTLF is to reselect a new ML model or retrain the existing ML model that provided to the AnLF when it determines ML model degradation by either:

- MTLF determining ML model degradation by collecting new test data (including input data, ground truth data and the corresponding inference) and testing the ML model accuracy. MTLF can compute accuracy by comparing the predictions and the corresponding ground truth data.

NOTE 1: Input data is the necessary data which is collected by AnLF to perform inference to generate prediction and the ground truth data is the actual measured data which corresponds toa prediction.

- MTLF can collect data for monitoring purposes from AnLF, ADRF or other NF. When ADRF is used, the MTLF can retrieve the data by specifying in the request the DataSetTag.

NOTE 2: The DataSetTag is defined from the conclusions of KI#4.

- MTLF collects unified feedback (i.e. "hit" or "not hit" indication) of the Analytics ID forwarded by the AnLF from the analytics consumer .

- MTLF subscribes to AnLF, that is registered in MTLF with its accuracy monitoring for a model provided by that MTLF, for getting notifications of the accuracy degradation of the analytics generated by the model, where the AnLF determines accuracy information based on any of the following:

- Comparing predictions and its corresponding ground truth data.

NOTE 3: The ground truth data and the corresponding prediction is to be defined per Analytics ID.

- Getting unified feedbacks (“hit” or “not hit” indication) of an Analytics ID from the prediction analytics consumeraccording to the difference between the KPIs related to the performance without/before taking the analytic into account and the one with taking the analytic into account..

NOTE 4: The analytics consumer can provide the unified feedback of an Analytics ID in accuracy feedback notification message by implicit subscription in Nnwdaf\_AnalyticsInfo\_Request Response or Nnwdaf\_AnalyticsSubscription\_Subscribe Response.

- Comparing changes in internal configuration for the analytics ID generation (e.g., data collection parameters).

- Previous existent records of analytics accuracy information.

- AnLF/MTLF can evaluate the quality of the data from the 3rd party data sources for input data selection.

**Triggers of performance check:**

- MTLF with accuracy checking capability of ML models can trigger the analytics accuracy checking based on its internal logic or configuration which may require to subscribe events, i.e., a change in the policy and/or a change in the subscription data for Target of ML Model Reporting, etc

- When requesting an ML model via the MLModelProvision service, the AnLF can specify in the request the additional parameters indicating the need for ML model accuracy check.

- When MTLF provides an ML model to an AnLF, the MTLF requests/subscribes AnLF to determine accuracy of the analytics generated from that model by comparing predictions and its corresponding ground truth data, if the AnLF indicates it can provide accuracy feedback.

- An analytics consumer may request or subscribe to accuracy information about Analytics ID(s) from the AnLF with the performance checking capabilities. Accuracy information can be included in an accuracy report, scoped in the same way as Analytics requests are scoped, i.e., per Analytics ID, for a specific area, slice, (group of) UEs, in a given time window, etc. Such request or subscription triggers the monitoring and check of Analytics ID(s) and generation of analytics accuracy information.

**Actions after** accuracy **check:**

* When accuracy information includes an indication that the accuracy of the analytics does not meet the consumer’s requirements, the analytics consumer may stop using analytics for a period of time or obtain new analytics.

In addition, accuracy information may also include updated analytics for the provided analytics ID, if the updated analytics is able to be generated within the correction time period.

- When accuracy information includes indications for the NF to stop or pause the consumption of the analytics, the NF may unsubscribe to the analytics ID, or provide an indication to AnLF that it is pausing an existing subscription of the analytics ID. Once AnLF determines the accuracy of the analytics is improved to meet the consumer’s requirements for an analytics ID, the AnLF may notify the NF consumer with an indication for resuming consumption of analytics ID.

- NF consumers of Analytics ID(s) upon receiving an accuracy information from an AnLF may request a pause or resume of notification from existing subscriptions

**Other aspects:**

- In order to improve correctness of NWDAF Service Experience analytics, the AF may provide "Service Experience Contribution Weights" to the NWDAF as described in Solution #2.

- Providing Multiple ML models to AnLF may help improve Analytics accuracy. In this case, each ML model shall indicate the providing MTLF and is assigned a unique ML Model identifier (i.e. unique within a PLMN) by the providing MTLF.

NOTE 4: The structure and format of the ML Model identifier and its uniqueness are up to stage 3.

- When requesting an ML model via the MLModelProvision service, the AnLF can specify in the request the information about input data type to assist MTLF in the ML model selection.

Editor’s Note: It is FFS if data granularity also included in the request.

\* \* \* \* Second change \* \* \* \*

\* \* \* \* End of changes \* \* \* \*