TR-200055

Technical Report

5G Automotive Association; Working Group System Architecture and Solution Development; 5GS Enhancements for Providing Predictive QoS in C-V2X

Contents

Foreword

This Technical Specification has been produced by 5GAA.

The contents of the present document are subject to continuing work within the Working Groups (WG) and may change following formal WG approval. Should the WG modify the contents of the present document, it will be re-released by the WG with an identifying change of the consistent numbering that all WG meeting documents and files should follow (according to 5GAA Rules of Procedure):

x-nnzzzz

(1) This numbering system has six logical elements:

(a) x: A single letter corresponding to the working group:

where x =

T (Use Cases and Technical Requirements)

B (Business Models and Go-To-Market Strategies)

A (System Architecture and Solution Development)

S (Standards and Spectrum)

P (Evaluation, Testbed and Pilots)

(b) nn: Two digits to indicate the year. i.e. 16, 17, 18, etc.

(c) zzzz: Unique number of the document

(2) No provision is made for the use of revision numbers. Documents which are a revision of a previous version should indicate the document number of that previous version

(3) The file name of documents shall be the document number. For example, document S-160357 will be contained in file S-160357.doc

Introduction

It is widely recognised that 5G will help automotive industries to achieve the vision of connected and automated driving with enhanced safety. Managing Quality of Service (QoS) is one of the most critical and challenging aspects for connected and automated driving to be accepted in reality, as various pre-agreed QoS Key Performance Indicators (KPIs), such as throughput, latency, and packet delivery ratio may not be guaranteed at all times. Enabling notifications with QoS predictions to vehicle applications in the 5G system presents a way to tackle the issues that could be caused by a potential QoS degradation. Hence, it is possible to improve overall system reliability while enhancing the safety of connected and automated driving.

A study on Predictable QoS and E2E Network Slicing for Automotive Use Cases (WI 'NESQO' [1]) was performed in 5GAA WG2 during 2018. When this Work Item was closed it was concluded that further study would be needed, the results of which are now documented in this report. The main focus of this work is on further detailing aspects and mechanisms for making QoS predictions, and on application and network reactions to such predictions.

Briefly, this document is organised as follows: Sections 1, 2, and 3 give the scope of this study, references, definitions and abbreviations used throughout this document, respectively. Section 4 presents analysis for use cases that are foreseen to benefit from Predictive QoS. Section 5 presents methods and procedures for making QoS predictions on the network and application levels. It also includes an analysis of Prediction Function (PF) location and discussions on the delivery of QoS prediction notifications to V2X applications. Further, it includes discussions on reaction of applications to QoS prediction and network reactions to coverage change predictions. In addition, Section 5 provides suggestions to 3GPP on further improving the standardisation related to predictive QoS based on this study. Section 6 concludes the Technical Report.

1 Scope

The NESQO Work Item (WI) kicked off in February 2018 [1] with input from car makers, in order to address key automotive requirements [2] [3] on a selected set of identified use cases. The main objective of the study was to provide Predictive QoS for C-V2X in 5G, utilising technologies including End-to-End Network Slicing, Multi-Access Edge Computing (MEC), an evolved QoS framework and Machine Learning.

Upon conclusion of the NESQO WI, the need for a follow-up WI [4] was put forward; some of the topics already identified in the original WID were not fully covered in the NESQO TR [5] and needed further study. Additionally, new requirements had been identified, that were relevant to the agreed scope but not included in NESQO.

This document addresses the 5GAA WG2 Work Item 'Enhanced End-to-End Network Slicing and Predictive QoS' [1].

WG2 understands that the present document is updated at each WG2 meeting during the WI and captures the list and the description of technical features to be considered in this WI.

The scope of this document covers the deliverable 'WG2-00XX' as described in the Work Item Description (WID) under 'Expected Output and Time Scale'. [1]

2 References

The following documents contain provisions which, through reference in this text, constitute provisions of the present document.

- References are either specific (identified by date of publication, edition number, version number, etc.) or non-specific.
- For a specific reference, subsequent revisions do not apply.
- For a non-specific reference, the latest version applies.

[1] 5GAA_A-180281: 'Predictable QoS and E2E Network Slicing for Automotive Use Cases'

[2] 5GAA_B-180007: 'Agile Quality of Service Adaptation: An Enabler for Advanced Connected-Vehicle Applications'

[3] 5GAA_A-180002: 'Enhanced QoS and Network Coverage Provisioning and Predictability for C-V2X'

[4] 5GAA_A-190002: 'Enhanced End-to-End Network Slicing and Predictive QoS'

[5] 5GAA_A-190176: 'Architectural Enhancements for Providing QoS Predictability in C-V2X'

[6] 5GAA_A-170188: '5GAA V2X Terms and Definitions'

[7] 5GAA T-180205: 'Use-Case Description Tele-operated Driving'

[8] 5GAA T-180250: 'Use-Case Description High-Definition Map Collecting and Sharing'

[9] 5GAA T-180146: 'Use-Case Description In-Vehicle Entertainment'

[10] 3GPP TS 23.288 V16.2.0 (2019-12) 'Architecture Enhancements for 5G System (5GS) to Support Network Data Analytics Services (Release 16)'.

[11] Bontempi, G., Taieb, S. B., & Le Borgne, Y. A. (2013). 'Machine Learning Strategies for Time Series Forecasting'. B*usiness Intelligence* (pp. 62-77). Springer Berlin Heidelberg.

[12] Barth, D., Bellahsene, S., & Kloul, L. (2012, October). 'Combining Local and Global Profiles for Mobility Prediction in LTE Femtocells'. *Proceedings of the 15th ACM International Conference on Modelling, Analysis and Simulation of Wireless and Mobile Systems* (pp. 333-342). ACM.

[13] Samba, A., Busnel, Y., Blanc, A., Dooze, P., & Simon, G. (2017, May). 'Instantaneous Throughput Prediction in Cellular Networks: Which Information is Needed?' *Proceedings of IFIP/IEEE International Symposium on Integrated Network Management* (IM). IEEE, 2017.

[14] Samba, A., Busnel, Y., Blanc, A., Dooze, P., & Simon, G. (2018). 'Predicting File Downloading Time in Cellular Network: Large-Scale Analysis of Machine Learning Approaches'. *Computer Networks*, 145, 243-254.

[15] Jomrich, F., Fischer, F., Knapp, S., Meuser, T., Richerzhagen, B., Steinmeth, R. (2018, June). 'Enhanced Cellular Bandwidth Prediction for Highly Automated Driving'. *Smart Cities, Green Technologies, and Intelligent Transport Systems 7th International Conference*. Springer, 2018.

[16] RFC 8321: 'Alternate-Marking Method for Passive and Hybrid Performance Monitoring'.

[17] 3GPP TR 23.786 v16.1.0: 'Study on Architecture Enhancements for the Evolved Packet System (EPS) and the 5G System (5GS) to Support Advanced V2X Services (Release 16)', 2019.

[18] 3GPP TS 23.287 v16.1.0: 'Architecture Enhancements for 5G System (5GS) to Support Vehicle-to-Everything (V2X) Services (Release 16)', 2019.

[19] Google (2019, June). Android Telephony Android API. Link: https://developer.android.com/reference/android/telephony/package-summary

[20] Apple (2019, June). iOS Core Telephony API. Link: https://developer.apple.com/documentation/coretelephony

[21] 3GPP TS 23.286 V16.2.0 (2019-12): 'Application Layer Support for Vehicle-to-Everything (V2X) Services; Functional Architecture and Information Flows (Release 16)'.

[22] 3GPP TR 23.764 V0.3.0 (2019-12): 'Study on enhancements to application layer support for V2X services (Release 17)'.

[23] 5GAA_A-190038: 'Seamless Integration of Vehicles as IoT Devices Using LTE-M'.

[24] 3GPP TS 23.401 v16.3.0: 'General Packet Radio Service (GPRS) Enhancements for Evolved Universal Terrestrial Radio Access Network (E-UTRAN) Access', June 2019.

[25] 3GPP TS 36.306 v15.5.0: 'Evolved Universal Terrestrial Radio Access (E-UTRA); User Equipment (UE) Radio Access Capabilities (Release 15)', June 2019.

[26] 3GPP TS 36.331 v15.6.0: 'Evolved Universal Terrestrial Radio Access (E-UTRA); Radio Resource Control (RRC); Protocol Specification (Release 15)', June 2016.

[27] 3GPP TS 23.501 V16.3.0 (2019-12): 'System Architecture for the 5G System (5GS); Stage 2(Release 16)'.

[28] 3GPP TR 26.939: 'Guidelines on the Framework for Live Uplink Streaming (FLUS)'.

[29] IETF RFC 8298: 'Self-Clocked Rate Adaptation for Multimedia'.

[30] I. Johansson, S. Dadhich, U. Bodin, T. Jönsson (August 2018). 'Adaptive Video with SCReAM over LTE for Remote-Operated Working Machines, Wireless Communications and Mobile Computing'.

[31] 5GAA A-190288: 'UE-based QoS prediction complementing network based prediction'.

[32] 3GPP TS 28.554 V16.2.0 (2019-09): 'Management and orchestration; 5G End-to-End Key Performance Indicators (KPI) (Release 16)'.

[33] 5GAA T-190179 5GAA: 'Work Item Draft Description – Predictive QoS and V2X Service Adaptation (PRESA)'.

[34] 3GPP TS 23.434 V16.2.0 (2019-12): 'Service Enabler Architecture Layer for Verticals (SEAL); Functional Architecture and Information Flows (Release 16)'.

[35] 5GAA A-190250: 'V2X application Layer Reference Architecture'.

[36] 3GPP TS 23.682 V16.5.0 (2019-12): 'Architecture Enhancements to Facilitate Communications with Packet Data Networks and Applications (Release 16)'.

3 Definitions and Abbreviations

3.1 Definitions

Definitions of terms in this document are elaborated in 5GAA TR A-170188 [6].

3.2 Abbreviations

For the purposes of the present document, the following abbreviations apply:

3GPP	Third Generation Partnership Project
5GS	5G System
5QI	5G Quality of Service Indicator
AF	Application Function
AMF	Access Management Function
API	Application Programming Interface
ARIMA	Autoregressive Integrated Moving Average
ARMA	Autoregressive Moving Average
AS	Application Server
CDF	Cumulative Distribution Function
CE	Coverage Enhancement
CQI	Channel Quality Indicator
C-V2X	Cellular Vehicle-to-Everything
E2E	End-to-End
FLUS	Framework for Live Uplink Streaming
GBR	Guaranteed Bit Rate
GFBR	Guaranteed Flow Bit Rate
HD	High Definition
HSS	Home Subscriber Server
IQN	In-advance Quality of Service Notification
IVE	In-Vehicle Entertainment
KPI	Key Performance Indicator
LTE-M	Long Term Evolution – Machine-Type Communication
MBR	Maximum Bit Rate
MEC	Multi-Access Edge Computing
MDT	Minimisation of Drive Tests
MFBR	Maximum Flow Bit Rate
MNO	Mobile Network Operator
NAS	Non-Access Stratum
NEF	Network Exposure Function
NG-RAN	Next Generation Radio Access Network
NMS	Network Management System
NWDAF	Network Data Analytics Function
OAM	Operations, Administration and Management
OEM	Original Equipment Manufacturer
OSI	Open Systems Interconnection
OTT	Over-The-Top
PCF	Policy Control Function
PDB	Packet Delay Budget
PER	Packet Error Rate

PF	Prediction Function
QoS	Quality of Service
RRC	Radio Resource Control
RSRQ	Reference Signal Received Quality
SARIMA	Seasonal Autoregressive Integrated Moving Average
SBI	Service Based Interface
SCReAM	Self-Clocked Rate Adaptation for Multimedia
SMF	Session Management Function
SNR	Signal to Noise Ratio
S-NSSAI	Single – Network Slice Selection Assistance Information
SPID	Subscriber Profile Identifier
ToD	Tele-operated Driving
UE	User Equipment
UPF	User Plane Function
V2X	Vehicle-to-Everything
VAE	V2X application Enabler
VAR	Vector Auto Regression
WI	Work Item
WID	Work Item Description
WG	Work Group

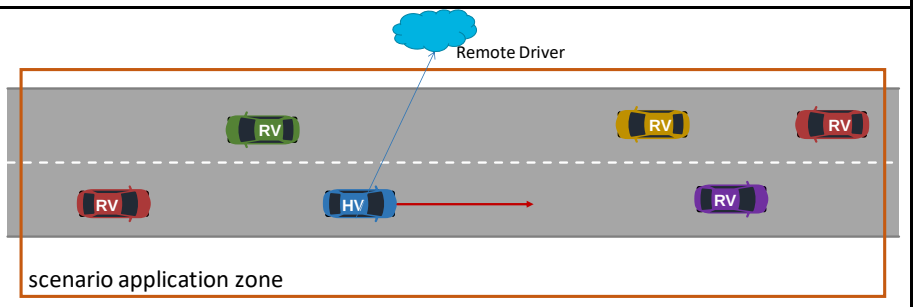
4 Use-Case Analysis

This section gives examples of use cases which are foreseen to benefit from Predictive QoS. Focus in the description is on the events flow when QoS predictions are made.

4.1 Tele-Operated Driving

A description of the Tele-Operated Driving use case is given in [7]. An example of an event flow and an example of QoS change values (QoS deterioration and associated time values), when QoS prediction is made, are given in Table 4.1-1.

Table 4.1-1: Prediction-centric description of ToD use case

Use-Case Name	Tele-operated Driving (ToD).
Vehicle Roles	Host Vehicle (HV) represents the remotely driven vehicle. Remote Vehicles (RV).
Other Actors' Roles	The Remote Driver (human or machine) undertakes to drive remotely the HV.
Illustrations	
Pre-conditions	- Vehicle HV is connected to 5G. - Command messages are sent from Remote Driver to HV every 20 ms. - Vehicle HV is equipped with four HD cameras, each with a data rate of 15-29 Mbps, plus 4 Mbps for object data. - ToD service has four adaptation states with the following UL data rates: < 60 Mbps, 60-88 Mbps, 88-120 Mbps, and > 120 Mbps.
Main Event Flow	• HV is driven remotely, Remote Driver is a machine: - The Remote Driver receives road conditions (e.g. obstacles) and status information of neighbouring RVs (e.g. location, speed, dynamics, etc.) derived, for instance, by HV's sensors, status information of the HV (e.g. speed, location), and traffic conditions. - Based on the information received, the Remote Driver builds the model of surroundings (i.e. awareness of the HV environment) and, taking into account the destination point, selects the trajectory and manoeuvre instructions. - The HV receives from the Remote Driver trajectory and/or manoeuvre instructions and executes them, according to HV's on-board security checks. - Feedback is provided to the Remote Driver in parallel with the execution of the manoeuvre. • HV is driven remotely, with Service Level Latency for command messages from Remote Driver to HV less than 20 ms.

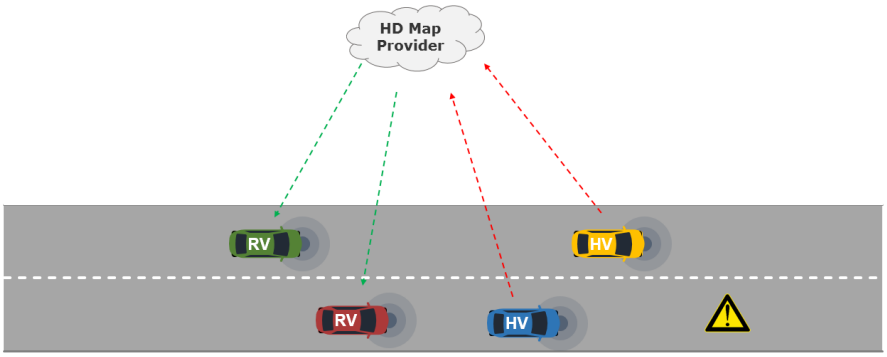
	• PF predicts QoS (latency) may degrade to the range of 20-30 ms in 30 s. • PF sends QoS prediction notification to Remote Driver. • Remote Driver reduces the speed of the HV, if it can still continue to safely drive the HV despite the higher latency (depending, for example on speed and surrounding environment), or alerts the human driver in the HV to resume control, or takes the HV to a safe stop.
Alternative Event Flow	• HV is driven remotely, Remote Driver is a human: - The Remote Driver receives high-quality video streams (e.g. to identify road conditions, neighbouring RVs) and the HV's status information (e.g. speed, location). - Based on the information received, the Remote Driver builds its situation awareness and, taking into account the destination point, selects the trajectory and manoeuvre instructions. - The HV receives from the Remote Driver trajectory and/or the manoeuvre instructions and executes them, according to HV's on-board security checks. - Feedback is provided to the Remote Driver in parallel with the execution of the manoeuvre. • HV is driven remotely with UL data rate for video more than 120 Mbps, and a latency of less than 20 ms. • PF predicts QoS (UL data rate) may drop to the range 88-120 Mbps in 20 s. • PF sends QoS prediction notification to Remote Driver. • Remote Driver reduces the speed of the HV such that it can continue to safely drive the HV despite lower video quality (resulting from lower UL data rate).
Post-conditions	The HV is driven safely with reduced speed, or if needed taken to a safe stop.

4.2 High-Definition Map Collecting and Sharing

A description of High-Definition Map Collecting and Sharing is given in [8]. An example of an event flow where a QoS prediction is made; when an automated driving vehicle is collecting, sharing and using HD map information to make optimal driving decisions and an example of QoS change values (QoS deterioration and associated time values) is given in Table 4.2-1.

Table 4.2-1: Prediction-centric description of HD Map Collecting/Sharing use case

Use-Case Name	HD map collecting and sharing.
Vehicle Roles	HV collects information about its surrounding using its own sensor devices, and shares the information with a HD map provider. RV receives HD map; each HV can also have the role of RV.

Other Actors' Roles	HD map provider that collects sensor information from HVs – and optionally also by road and roadside infrastructure – to build the HD map.
Illustrations	
Pre-conditions	The vehicles are connected to 5G and can make optimal driving decisions based on an up-to-date, precise, and reliable vision of the environment by using the real-time, highly accurate HD map. The HVs are equipped with sensors and they can share sensor information. The HD map provider can collect (and merge) sensor information from different sources to build the HD map in a fast and reliable manner. Service Level Latency of 100 ms is needed for HD map's end-to-end, real-time performance [6].
Main Event Flow	• HV1 (blue) is driven with automation level 4 or 5, with Service Level Latency of 20 ms. • PF predicts QoS (latency) may degrade to > 100 ms in 20 s. • PF sends QoS prediction notification to HV1 and optionally also to HD Map App Server. • HV1 steps down automation level and/or resumes manual operation. • HD Map App Server sends warning message to affected vehicles in the area.
Post-conditions	HV1 is driven safely at a lower level of automation. Potentially other vehicles may receive IQN and implement proper reaction.

4.3 In-Vehicle Entertainment (IVE)

A description of In-Vehicle Entertainment use case is given in [9]. An example of an event flow and an example of QoS change values (QoS deterioration and associated time values), when QoS prediction is made, is given in Table 4.3-1.

Table 4.3-1: Prediction-centric description of IVE use case

Use-Case Name	In-Vehicle Entertainment (IVE).
User Story	High-definition (HD) Content Delivery.
Short Description	The use case concerns the delivery of entertainment content to passengers within a moving vehicle. It is applicable to both automated

	and non-automated vehicles, where in the latter drivers are restricted in the content they are allowed to consume.
Actors	Host Vehicle, HV owner, operator or manager, passengers, service providers (e.g. wireless network operators, road operators, streaming and gaming services, a combination of them, and others).
Vehicle Roles	Host Vehicle (HV) is the vehicle where the passengers consume the content.
Pre-conditions	• Vehicle HV is connected to 5G. • The communication link for the HD content does not disrupt the communication link for other use cases involving safety and other mission critical services.
Main Event Flow	• Two passengers individually choose which HD content they are interested in before or after entering the car, each on HD 4k video streams (low-end service for cars [5]). Each stream needs an estimated 50 Mbps, and 150 ms latency. • HV is driven with DL data rate for video more than 100 Mbps. • PF predicts QoS (DL data rate) may drop to the interval 40-50 Mbps in 20 s. • PF sends QoS prediction notification to HV. • Application adaptation is made so that resolution for video is reduced, corresponding to a data rate of 2*20 Mbps. • When DL data rate drops, application adaptation has already taken place and there is no interruption in the video streaming service.
Post-conditions	Both passengers continue their video consumption, at lower resolution.

5 System Analysis and Architecture Enhancements for In-Advance Notifications Including Predictive QoS

This section contains an investigation into making QoS predictions, on network level and application level respectively. It also includes an analysis of Prediction Function location, as well as an analysis of application and network reaction to QoS predictions.

5.1 Making Network-Level QoS Prediction

This section contains an investigation into different strategies for the generation of prediction notifications, also considering accuracy information. It also contains an analysis of predictions enhanced with predicted network event information.

5.1.1 Generation of Prediction Notification Considering Accuracy Information

The generation of a prediction notification is obviously influenced by the predicted behaviour of the KPI under consideration and the associated threshold (or condition) for generating the prediction. In addition, another aspect to consider is the accuracy associated to a predicted KPI, which impacts the

probability that the actual KPI will behave as predicted and, in particular, as notified to the prediction consumer. Generally speaking, a prediction with associated high accuracy means that there is a high probability that the KPI will actually turn out as predicted and, consequently, that the generated notification represents with adequate precision the expected future outcome of the KPI. On the other side, a prediction with a low accuracy means that there is a certain probability (the lower the accuracy, the higher the probability) that there will be a deviation of the actual KPI compared to what has been predicted and notified to the consumer. From this point of view, consumers of prediction notifications may react in a different way to predictions with different accuracy levels. For instance, consumers may trigger more conservative reactions in case of a prediction with low accuracy to compensate for the probability of a false prediction, while reactions may be less conservative in case the prediction has associated higher accuracy. The information about the accuracy associated with the prediction task could be directly exploited in the process of generating prediction notifications. This is the focus of the remainder of the section.

Consider the example where a certain V2X service is on-going (i.e. it is running and relevant consumers have already subscribed to prediction notifications). It is assumed that prediction information comprises notifications received in advance. In this example, the final aim for the 'prediction consumer' is to be aware if the KPI is predicted to be below a certain threshold, e.g., in a similar way as considered in 'QoS Sustainability Analytics' by 3GPP TS 23.288). From this point of view, the notification can focus either on providing prediction notifications covering the whole planned journey or on providing prediction notifications before the vehicle approaches areas where the KPI is predicted below the threshold (or a combination of the two). As an illustration, Figure 5.1-1 depicts the prediction of a certain target KPI. In particular, it shows with a blue continuous line the *predicted value*, i.e., the value that the prediction algorithm predicts as the one that will be experienced. With an orange dotted line the *lower/upper bounds* of the predicted KPI are shown. The line is generated considering the accuracy of the prediction, i.e. the *lower/upper bound* increases if the accuracy of the prediction decreases[1]. As mentioned above, the need from the prediction consumer is to be informed when the KPI is predicted to be below a certain threshold (red dotted line). In Figure 5.1-1, two examples are depicted where accuracy level changes over time. In the first example, the predicted value of the KPI is expected to be below the threshold. In the second example, the predicted value is expected to be above the threshold, but the fact that there is a lower accuracy may bring the actual value below the threshold.

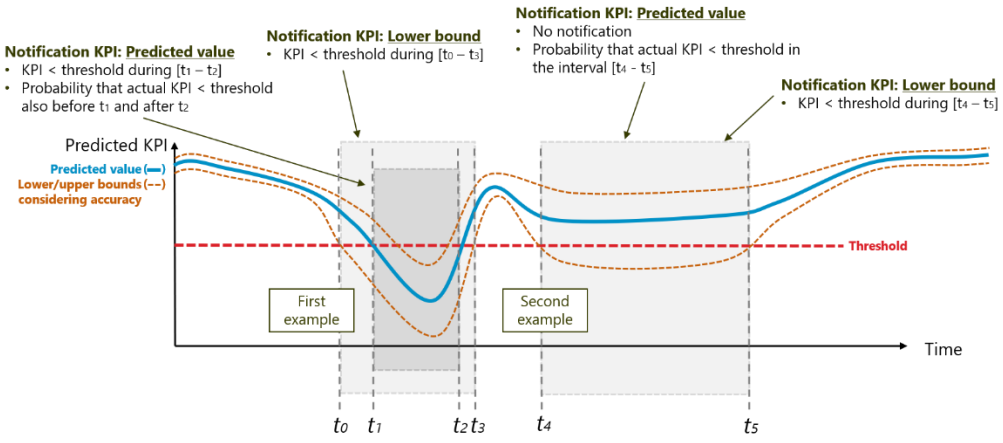


Figure 5.1-1: Examples of different notifications triggered considering different strategies for the generation of prediction, where the task is to predict whether a certain KPI will be below a pre-defined threshold

[1] This is to illustrate that lower accuracy implies that the actual value of the KPI that will be experienced may fall in a larger range of values.

We firstly consider the case when the *prediction is generated with the aim of notifying the consumer about the predicted value* (blue continuous line in Figure 5.1-1). In this case, the prediction is generated if the predicted value of the KPI is expected to be below the threshold. This implies that the prediction notification includes the predicted value and its accuracy. As a consequence, from Figure 5.1-1, we have:

- *First example*. The notification of QoS prediction would be generated to indicate that the *predicted* KPI is below the threshold during time interval [t_1-t_2]. Nevertheless, as highlighted in the drawing, there is lower prediction accuracy in proximity of time t_1. Therefore, the actual KPI may fall below the threshold also before time t_1 (the same holds for the time after t_2). If this happens, it will then generate a false prediction. A possible way to mitigate this is for the prediction notification to indicate that: during [t_0-t_1], the predicted value of the KPI is above the threshold, but with a certain probability (derived from the accuracy of prediction) that it may fall below the threshold; during [t_1-t_2], the predicted value of the KPI is below the threshold; and during [t_2-t_3], the predicted value of the KPI is above the threshold, but with a certain probability (derived from the accuracy of prediction) that it may fall below the threshold.
- *Second example*. In its basic form, the prediction would not generate any notification as the *predicted value* of the KPI is above the threshold. Nevertheless, as highlighted in the drawing, there is lower prediction accuracy during time interval [t_4-t_5], which means the actual KPI may fall below the threshold. If this happens, it will then generate a false prediction. A possible way to mitigate this is that the prediction notification indicates that during time interval [t_4-t_5] the predicted value of the KPI is above the threshold, but with a certain probability (derived from the accuracy of prediction) that it may fall below the threshold.

The discussion above can be extended considering the case when the *prediction is generated with the aim of notifying the consumer about the predicted lower bound of the predicted KPI,* which can be derived from *the accuracy of prediction* (orange dotted line in Figure 5.1-1. In this case, the prediction is generated if the *lower bound* of the calculated KPI – considering the predicted value and accuracy of prediction – is expected to be below the threshold. This implies that the prediction notification includes the predicted *lower bound* **and** its accuracy. As a consequence, from Figure 5.1-1, we have:

- *First example*. The notification of QoS prediction would be generated to indicate that the predicted lower bound KPI is below the threshold during time interval [t_0-t_3].
- *Second example*. The notification of QoS prediction would be generated to indicate that the predicted lower bound KPI is below the threshold during time interval [t_4-t_5].

A possible side effect of this approach is related to the fact that the predicted lower bound might be much lower than the predicted value of the KPI. This may cause inefficiencies at consumer side, e.g., taking too conservative decisions based on the predicted lower bound performance even though the experienced ones will be better.

Additional cases may also consider that prediction notification includes both the predicted value and the predicted lower bound. Another case is when the prediction notification includes the *predicted lower/upper bounds* of the KPI under analysis. The actual tuning of prediction notifications will ultimately depend on the needs of the associated consumer, and this will depend on how the consumer is designed to react to notification of QoS prediction.

5.1.2 Prediction Enhanced with Predicted Network Event Information

The prediction is influenced by several factors, including the presence of *network events* which may happen during the time window to be predicted and may impact the behaviour of KPIs (bit rate, latency,

etc.). This is of particular relevance when considering Prediction Functions (PF) located within the domain of a network operator. Network events, such as handover, which is of course of high relevance for vehicles, may have an impact because the prediction of a KPI during the handover is affected by handover-specific features. Indeed, in addition to other features which are more typical of prediction tasks, e.g., cell load at source and target cells, aspects such as the time (and/or place) when the handover is triggered, handover completion time, temporary loss of user plane connectivity for a certain amount of ms, etc., may impact the prediction as well. Therefore, the knowledge of features associated with network events impacting prediction tasks may be considered as well to improve the prediction. Considering that prediction tasks focus on future time windows, then also in this case the focus would be on *predicted network event*. For example, specific network events such as handover could be monitored, with the relevant information collected and potentially predicted. Continuing the handover example, that could mean collecting information about the rate of handover events in a certain location, timestamps of when handover is triggered and completed, the handover duration, and statistics on target KPI behaviour during handover, which could all be used to generate predicted information associated with the network event. This gives availability of detailed information to describe the predicted network event, information that could be exploited by prediction algorithms at the PF.

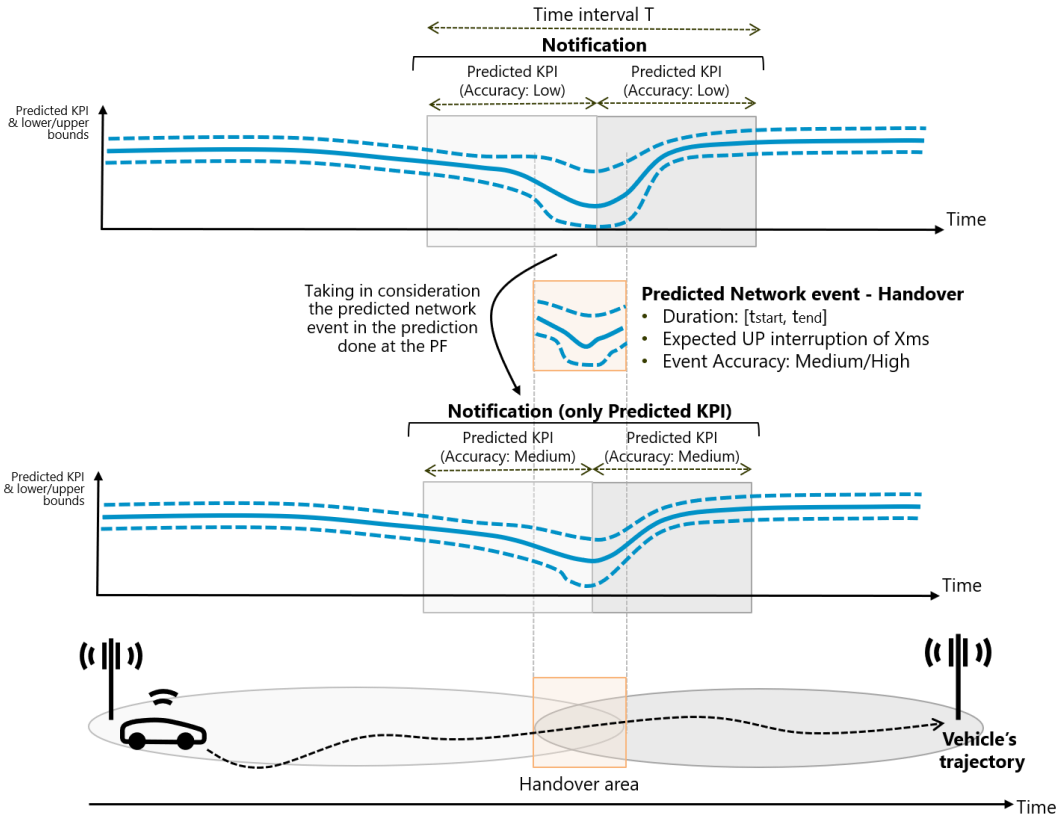


Figure 5.1-2: Example of prediction task enhanced with information about predicted network event which is used by PF to improve prediction

What we have discussed above can be further elaborated on with an example, which assumes PF is located within operator domain. Figure 5.1-2 depicts the case when a prediction task at the PF oversees the predicting of a certain KPI. In Figure 5.1-2, the dotted lines are used as visual representation of lower/upper bounds of the predicted KPI, which can be considered to be associated with prediction accuracy (i.e. the higher the lower/upper bounds with respect to the predicted KPI, the lower the accuracy). It is assumed that information about expected vehicle trajectory is available at the PF. The

network event under consideration in this example is a handover. In Figure 5.1-2, the area where the handover is expected to happen during vehicle's journey is highlighted.

An improvement to discuss is when the PF exploits information about predicted network event to improve its prediction. In the top drawing of Figure 5.1-2, the prediction generated by the PF shows that the predicted KPI, within the time window of the handover, has large lower/upper bounds as representation of low accuracy of prediction, due to the presence, as an example, of the handover event which influences the accuracy of prediction as discussed above. In Figure 5.1-2 it is considered that the PF can exploit information about the predicted network event (i.e. handover). Such information could be considered as an *event description*, including fields such as event type (i.e. handover in this example), event duration (e.g., in the form of t_{start}-t_{end}), event-specific information (e.g., expected user plane interruption of X ms), prediction accuracy (or event probability), KPI behaviour during the event, etc. The information included in the event description could be used by the PF in an implementation-specific manner to improve its prediction. For instance, the PF may take into consideration the KPI behavior during the predicted event included in the event description and its associated event duration information as inputs to predict the KPI behaviour during the overall prediction time interval (T). Another example is that the PF may modify its behaviour by adapting its prediction windows. Indeed, in case of predicted network event, the time intervals considered in the KPI prediction are modified by 'excluding' the time interval when the handover is predicted to happen. In this case, the information in the event description is used by the PF to generate a prediction for the time interval where the prediction is expected to happen. This could help to improve prediction before and *after* the predicted network event and the availability of event description at the PF would allow to generate an accurate prediction covering the time interval of predicted network event. As mentioned above, whether a network event impacts a prediction task depends on the PF's implementation and also on the KPI(s) to be predicted (i.e. different network events may affect the various KPIs in different ways). In this example, the PF generates a notification towards relevant consumers, including information about predicted KPIs and associated accuracy (i.e. without extending the type of information included in the notification). In this case, the information about a predicted network event is used by the PF to improve the generation process of the prediction of the KPI under consideration, and to generate a prediction notification which captures the impact the predicted network event has on predicted KPIs. In this way, the PF can provide more accurate predictions to the relevant consumers.

Other possibilities in exploiting information generated by the PF as a result of the prediction of network events need to be further evaluated. For instance, further analyses may investigate the aspect of enhancing the information of the QoS prediction notification by including information generated by the PF as a result of the prediction of network events (e.g. handover). This information may improve the information available to the consumer concerning expected network performance, and may be used by the consumer/application in several ways. For instance, the information generated by the PF as a result of the prediction of network events such as handover, may generate a different type of reaction at consumer side compared to a prediction notification without information associated with predicted network events. Especially, more appropriate and customised application reactions may be triggered that are not perceived as harsh as reacting to a false prediction, where the spectrum of possible reactions could range, e.g., from lowering vehicle speed to a take-over request to the driver. From this point of view, further analyses are required to understand firstly if the information generated by the PF as a result of the prediction of network events is beneficial for the adaptation process of the associated consumer or if the information about the impact that predicted network event has on predicted KPIs is enough (for instance, the information about a predicted network event such as handover with a UP interruption of X ms could be delivered to the consumer as prediction of the KPI UP availability). From this point of view, it is for further studies to evaluate the content and format of this information generated by the PF as a result of the prediction of network events. Secondly, further analyses are required to understand which network events are relevant for the consumer to be notified, and under which circumstances such network events are relevant. These two points obviously depend on

application's implementation (i.e. only some applications may benefit from network event prediction and not all network events may impact the application). Finally, it should be considered that the exposure of information generated by the PF as a result of the prediction of network events depends on agreements between the PF provider and the associated consumer. This aspect is of particular relevance when considering network operators as PF providers and the fact that predicted network event may represent network internal information whose exposure requires further discussions.

5.1.3 Threshold Options for Potential QoS Change Notifications

This section describes implementation aspects aiming at reducing the number of notifications that are triggered to the consumers, as well as the amount of data contained in the actual notification.

5.1.3.1 Reporting Threshold to Limit Notifications Overhead

In accordance with the PQoS concept [5], a prediction notification is triggered each time when the reporting threshold(s) is crossed by the predicted QoS KPIs as defined in TS 23.288 [10]. In case when the predicted KPI(s) is oscillating close to the threshold(s) by the expected QoS KPIs many notifications not relevant to the consumer are generated, as illustrated in Figure 5.1-3 by the time intervals marked in grey.

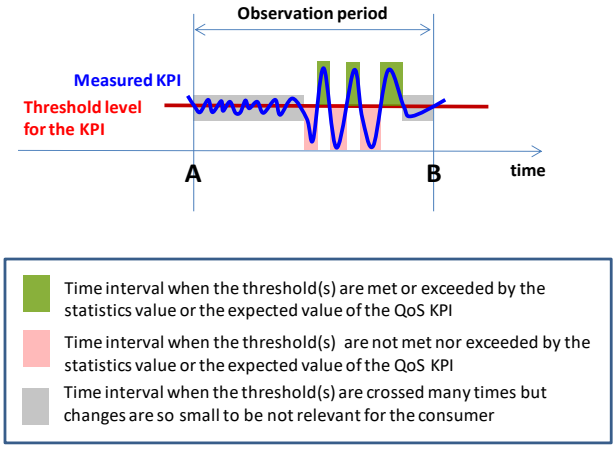


Figure 5.1-3: 'Unnecessary' applicable time interval leading to large signalling overhead

Note that notifications may be sent in both directions – QoS degradation (i.e. critical direction) and QoS improvement (i.e. non-critical direction) – which can lead to large signalling flooding.

The related countermeasure to avoid this overhead is illustrated in Figure 5.1-4.

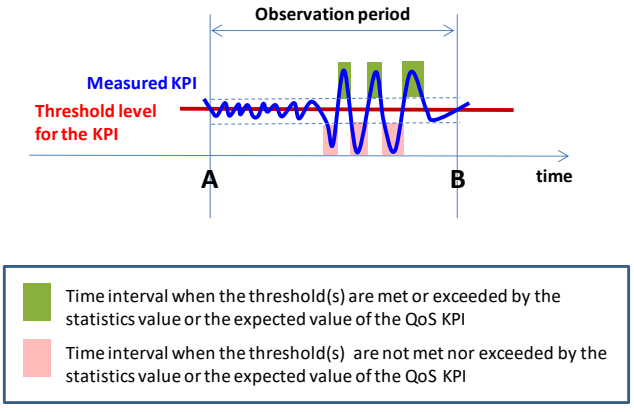


Figure 5.1-4: Relevant applicable time intervals

As it can be seen from Figure 5.1-4, only the green and red areas are potentially relevant for the consumer. When there is a small difference between the threshold and the QoS KPI no information should be added in the analytics output provided to the consumer. This requires that the request or subscription from the NF Consumer includes not only a level but also a threshold minimum delta or hysteresis.

Threshold minimum delta presents the customer's acceptable deviation from a threshold without triggering reporting, to limit the amount of signalling caused by the notifications when the predicted KPI is oscillating but still close to the threshold.

At the same time, it is reasonable to use the threshold minimum delta (hysteresis) only in the non-critical direction, i.e. when the QoS change is improving the situation. That is, the reporting threshold is set for the critical direction as the corresponding notification is related to a necessary reaction at application layer, while a notification about an improvement of the situation could be omitted to reduce the amount of signalling when the predicted KPI is too close to the threshold.

As a result, for threshold-based reporting (e.g. a notification is triggered when a reporting threshold is crossed) the amount of notifications can be reduced with the help of hysteresis (delta parameters) applicable for the non-critical direction.

5.1.3.2 Threshold to Restrict the Value Range of the Output Data in Notifications

In addition to threshold-based reporting where notification happens directly after the subscription (i.e. with the initial state and whenever the state changes, the consumer may want to have periodic notifications about the state when the next point in time (according to the periodicity) is reached. The periodic notifications, i.e. reporting over a certain interval of time is defined as the number of notifications per time unit (e.g. 1 notification per minute).

For periodic reporting, the whole output information (analytics) about predicted KPI(s) is included in the potential QoS change notification payload each time a notification is sent. This could lead to large amount of data in the notification message. But, the consumer may only be interested in periodic reports when the predicted KPIs are within a certain interval of values, i.e. to have only a part of the full set of analytics to facilitate processing of this information.

To restrict the value range of the output data in the notification message in a case of periodic reporting a start-stop approach can be applied. An illustration of the start-stop approach is shown in Figure 5.1-5.

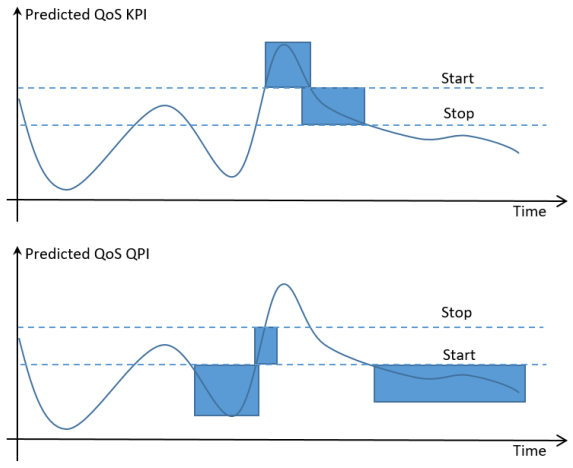


Figure 5.1-5: Start-stop approach to restrict value range of output information

The start-stop approach can be defined as follows.

If the start condition is higher than the stop condition then the notification includes the outcome information (analytics) when the level of predicted QoS KPI(s) goes above the 'start condition' and ends when the level of predicted QoS KPI(s) goes below the 'stop condition'.

Correspondingly, if the start condition is lower than the stop condition then the notification includes the outcome information when the level of predicted QoS KPI(s) goes below the start condition and ends when the level of predicted QoS KPI(s) goes above the stop condition.

As a result, for periodic or time-based reporting (i.e. a notification per time unit) the value range of the output information can be restricted by start-stop conditions related to a certain interval of values of predicted QoS KPI(s) which are relevant for the consumer.

5.1.4 Making UE-based QoS Prediction

User Equipment or UE-based prediction could be a complementary solution to the network-based solution. The UE based QoS prediction as a complementary solution to network based solution could alternatively map entities, such as In-advance QoS Notification (IQN) Producer, IQN Consumer, and Prediction Function (PF) in the UE, since the QoS prediction would be generated and consumed within the UE itself. In principal, the UE may rely on the network-based prediction for long-term time horizons and, in addition, it may rely on the UE-based prediction for short-term time horizons. In addition, for future scenarios a UE-based QoS prediction could be considered also for sidelink communication between UEs in out-of-coverage scenarios. UE-based predictions could be based on existing and new measurements collected in the UE. Such measurements can, for example, target the quality of received signals, the result of a procedure or the time consumption of a procedure. The UE external information (e.g. received from the network) could be applicable to setup and configuration of the prediction function within the UE. An application in the UE could use the information on predicted QoS generated by the UE, to adapt and take necessary measures (e.g. fall back to a pre-defined 'safe state').

A UE-based solution could, to a large extent, reuse collected information in the UE as input to the PF. As a part of the setup and configuration of the input to the PF the network can configure the UE with thresholds for the UE internal measurements. The thresholds could be tuned according to the requested – application specific – QoS and the capabilities of the network, as well as the UE. The format of the input to the PF could be actual measurement values of well-defined metrics or it can be indications. An indication could be based on the results of one or more measurements and triggered according to the network configured thresholds. It is worth noting that the UE-based prediction is mainly for the UE itself. Additionally, one UE's prediction, if applicable, may be shared/sent to the other UEs and/or to the network (5GS or AF). The details on what information is to be shared, and how it is shared and treated are topics for further studies.

To conclude, a UE-based solution can offer benefits both as a complement to a network-based solution for short-term predictions and as stand-alone solution in out-of-coverage scenarios.

5.2 Making Application-Level E2E QoS Prediction

This section contains an investigation of the means and one possible option on how to make an accurate prediction of E2E QoS using Machine Learning techniques. The framework presented in this section contains building blocks that should lead to an accurate QoS prediction. Every model

constituting the framework relies on a state-of-the-art Machine Learning technique. The first part of this section presents the Machine Learning method categories that the proposed framework uses.

5.2.1 Machine Learning Methods

We focus on two Machine Learning method categories used by the prediction framework: Supervised Learning and Time Series Modelling.

5.2.1.1 Supervised Learning for Mapping Models

Supervised Learning is a Machine Learning task that enables to infer the functional link (known as a **mapping model**) between two sets of variables. It enables to build a mapping model from data (a collection of experiences). For example, with a database that keeps an inventory of throughput test results on a cellular network together with several parameters collected at the beginning of each test, such as radio parameters and cell load and performance, with Supervised Learning we can build a mapping model between all the parameters and the throughput results [5] [6].

5.2.1.2 Time Series Models

Also known as 'statistical forecasting' methods, the purpose of time series forecasting is to predict temporal information. It consists of predicting a variable Y_{t+1} from the series $(Y_t, Y_{t-1}, Y_{t-2}, \dots, Y_{t-n})$ and potentially another/other correlated series $(X_t, X_{t-1}, X_{t-2}, \dots, X_{t-n})$.

The methods commonly used are Autoregressive Moving Average (ARMA, ARIMA, SARIMA, VAR, etc.), however other methods exist, such as applying classical Supervised Learning to time series [11].

5.2.2 Prediction Framework

The framework proposal is composed of four building blocks, as presented in the Figure 5.2-1. Each is detailed in the following paragraphs. NOTE: other options may be considered FFS

Network performance predictions and state forecasting can be composed of a set of several network parts (RAN, Core, backhaul, transport, etc.). The current prediction framework is focusing more on the RAN level, but other networkperformance indicators could be considered in the future.

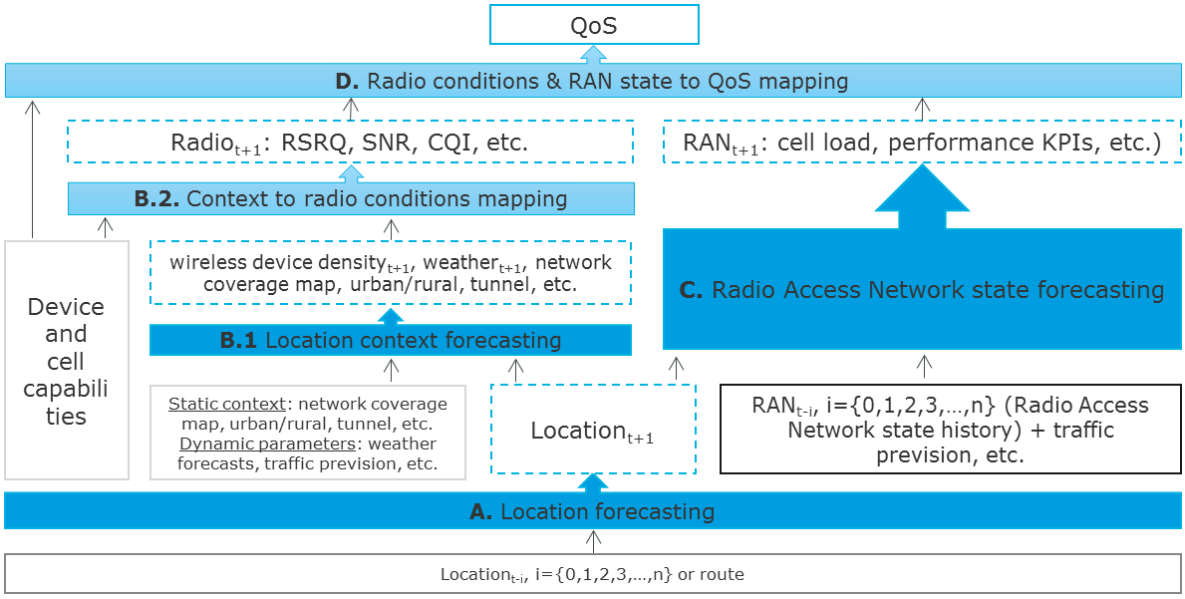


Figure 5.2-1: Prediction framework building blocks

5.2.2.1 UE Location Forecasting Based on Trajectory Models

The purpose is to use trajectory models [12] (using location history, speed, etc.) or the route directly provided by OEM's applications to forecast the position of the vehicle at time t+1.

5.2.2.2 Radio Conditions of a UE at a Precise Time and Location, Based on Data from Monitoring Tools

The radio conditions at time t depend on the factors of radio interference and noise related to the device location at time t (location context: wireless device density, weather, network coverage map, urban/rural, tunnel, etc.), and on the capacities of the device and the cell. Thus, a mapping model can be built – taking as input all of these parameters for a precise time and location – to infer the radio conditions. Otherwise, the task can be divided into two phases: first, a location context forecasting and then a context-to-radio condition mapping.

5.2.2.2.1 Location-Context Forecasting

Location-context forecasting consists in an inference from coverage maps, traffic prevision and crowd-sourcing data. In essence, it relies on structural and contextual parameters. Structural parameters are network coverage, the density of the area (rural or urban), the type of location (a tunnel, a highway, etc.).

5.2.2.2.2 Context-to-Radio Condition Mapping

A context-to-radio condition mapping model can be built taking into account the device and the cell capabilities.

Two types of data can alternatively help to build this mapping model:

- Geo-localised MDT (Minimisation of Drive Tests) data.
- Crowd-sourced location, context, radio condition data.

5.2.2.3 RAN Performance and State Forecasting

The Radio Access Network (RAN) performance and state forecasting is enabled by Times Series Modelling. It can rely on either RAN state history (from Network Management System) completed by traffic prevision or crowd-sourced data on radio access networks performance.

5.2.2.4 Mapping of Radio Conditions and RAN State to the Accessible QoS for UE

The model mapping radio conditions and RAN state to the accessible QoS uses Supervised Learning. It enables predictions of the QoS using as input the forecasted radio conditions and RAN state. An example of such a mapping model is outlined in earlier research [13] [14].

5.2.3 Model Learning

The learning phase of model composing the framework can respond to two scenarios detailed in the following paragraphs.

5.2.3.1 Learning scenario 1: Embedded Measurement Tool

This scenario provides the existence of a data-collection application embedded in vehicle devices. As illustrated in Figure 5.2-2, the embedded application performs radio condition and QoS measurements in real time and shares the results with the Machine Learning Engine hosting the Learning computation.

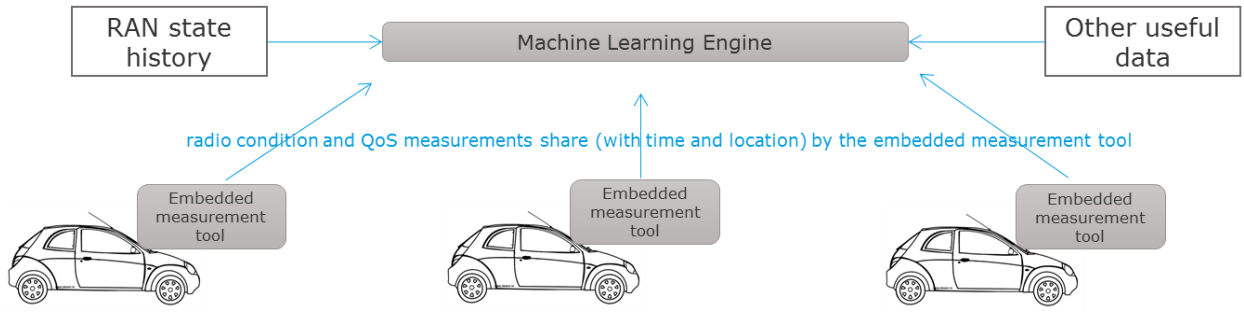


Figure 5.2-2: Embedded measurement tool scenario

5.2.3.2 Learning Scenario 2: NMS and Probe-Assisted Data Collection

This second scenario requires the activation on the cellular network of a feature similar to Minimisation of Drive Tests in order to collect radio condition data from a panel of vehicles, as illustrated in Figure 5.2-3. Corresponding QoS measurements can be retrieved using probes on actual mobile sessions.

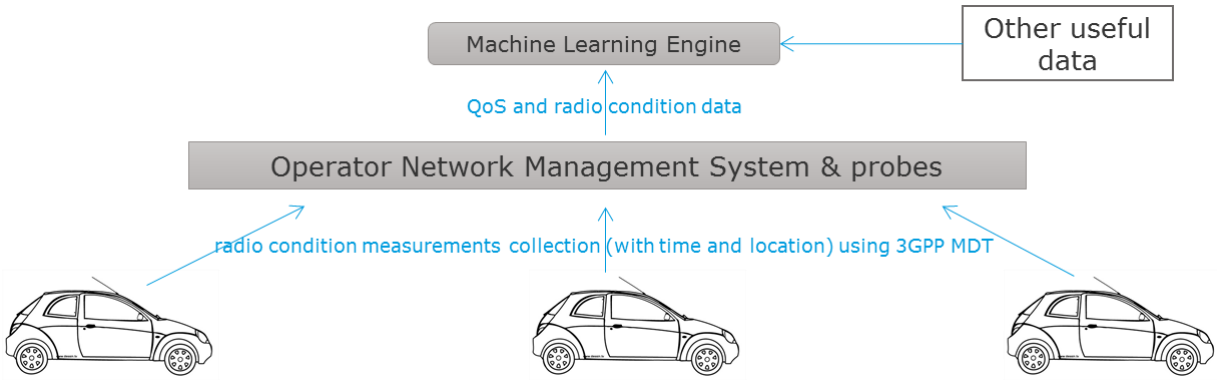


Figure 5.2-3: NMS and probe-assisted data collection scenario

5.2.4 Prediction Execution

Once the learning phase is accomplished, the prediction execution proceeds as illustrated in Figure 5.2-4.

First, the vehicle sends to the prediction framework host a QoS prediction request along with its forecasted location for time t+1. This forecasted location is used in the prediction framework, along with RAN data and other useful third-party data, by the location-context forecasting model (B.1). The B.1 model's results are used by the context-to-radio conditions mapping model (B.2). In parallel, the RAN and third-party data are used by the RAN forecasting model (C). Finally, the results of models C and B.2 are used as inputs for the QoS mapping model (D), to deliver the requested QoS forecast which is transmitted to the vehicle or open for authorised third-party applications.

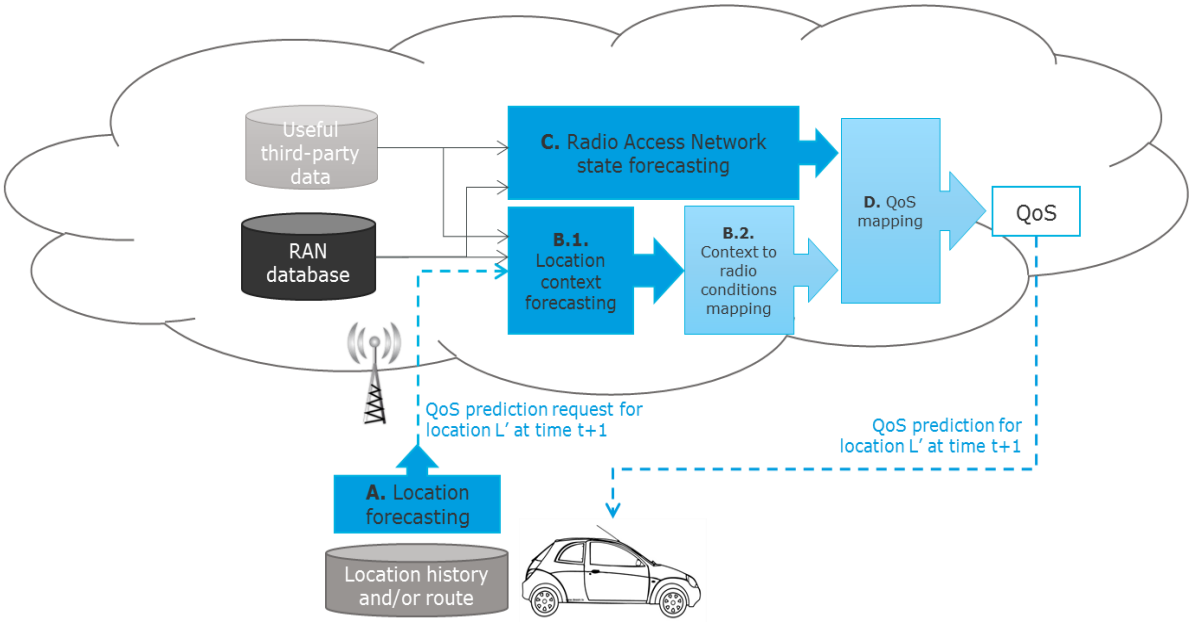


Figure 5.2-4: Prediction execution

5.3 Prediction Function Location

By design, relevant information to create predictions is distributed across components owned by the automotive and telecommunication industries. Based on the data accessible in each industry individually or shared across industries, the Prediction Function may achieve different results to meet the requirements per use case. This section provides a summary of two approaches: Over-The-Top or OTT[2]-based and Mobile Network Operator or MNO-based.

5.3.1 PF Deployed in OTT Domain

As an OEM or Car Tier 1 or generic third-party service provider, control over vehicle data creates the reasonable expectation that an own prediction function can be deployed and operated independent from the Mobile Network Operator. The main assets of this solution are the vehicle connectivity data gathered in the own or collective car footprint, where prediction can be then developed either (i) with a centralised approach where an OTT backend gathers measurements from the vehicle and generates the prediction then delivered to the vehicle, or (ii) with a decentralised approach where the vehicle directly generates predictions using its own measurements and possibly assisted by information from an OTT backend. If the prediction function is not relying on any MNO data and a high penetration rate of the information collection can be achieved, the prediction function may server multiple MNOs at once. Under this assumption fast geographical availability of Predictive QoS can be assumed.

Note that the same UE cannot generally survey multiple radio bands at the same time and not all UEs support local band-locking capabilities to force a certain set of bands to be used, which implies that the survey will be dependent on what the UE is being instructed to use from the network in terms of bands and technologies.

[2] OTT is used in the telecommunication industry as an 'overlay' solution not integrated in the telecommunication network. They usually operate on OSI layer higher than 3.

Unfortunately, the prediction created based only on vehicle data has limitations in terms of capacity and accuracy of the network.

Vehicle information can provide[3]:

- Detailed RF distribution maps for streets based on low-layer modem information in conjunction with highly accurate geo-location information.
- Device specific characteristics. This might also take into account specific UE capability configurations that could be OEM-specific, requiring ad-hoc solutions in the event prediction is offered by an OTT platform serving multiple OEMs.
- Passive data consumption information based on interface counters.
- (Optional) active bandwidth capacity information with active speed tests.

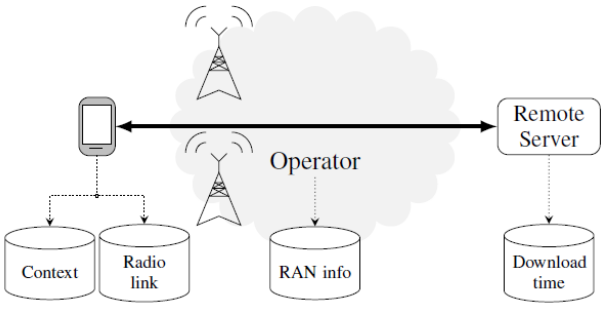
With modern Machine Learning technologies an OTT-player can create coverage maps for dedicated MNOs. Poor coverage regions, handover regions and call drops can be identified as a trend over a longer period of data collection. Due to slower propagation times and required statistical mass of probe data', network topology changes/upgrades and traffic pattern changes can only slowly be detected.

What cannot be considered with this data set are:

- Capacity information and resource utilisation of the serving cells.
- Specific traffic prioritisation enforced by the MNO (e.g. subscriber category, policy information).
- Mobility management optimisations for vehicular UE root causes, such as dynamic changes or outages of any network components[4].
- Network capabilities beyond the capabilities of the collecting UE.

It is anticipated that data from the OEM, combined with modern Machine Learning techniques, can over time provide statistically valid predictions for areas well covered by the OEM's data collection. This does not mean that any specific prediction at any specific point in time will be correct; only that a high sample of predictions will be statistically accurate. This might limit the availability of in-advance notifications about predictions of achievable network performance.

Current research approaches vary depending on the data sources available and analysed. Figure 5.3-1 and 5.3-2 illustrate research [13] [14] analysis of the impact of different training data sets on the prediction accuracy.



Data collection during measurement campaigns

[3] If very significant amounts of data are available.

[4] This may hide the information if the issue is structural or transient and also limit prediction capabilities when prediction of a specific QoS KPI requires the analysis of several cause KPIs that are not exported outside of MNO domain.

Figure 5.3-1: Study setup for data collection [3]

Figure 5.3-2 states clearly that a combination of all information sources collected from the UE and MNO for LTE throughput measurements achieve the highest accuracy. Studies like [15] have shown that UE-only based predictions are viable, but with constraints. Figure 5.3-2 shows for the example of LTE throughput measurements that higher accuracy of predictions has been achieved with network-based information sources.

		training	test	RF	
Predictors	# metrics	# entries	# entries	R^2	$\tilde{E}$
B	2	13,851	9,233	0.49	0.26
B + Radio	5	8,210	5,472	0.75	0.19
B + RAN	6	9,340	6,201	0.72	0.17
B + Context	5	10,710	7,140	0.59	0.24
B + Radio + RAN	9	5,872	3,903	0.87	0.11
B + Radio + Context	8	7,870	5,246	0.82	0.15
B + RAN + Context	9	7,320	4,863	0.86	0.13
B + Radio + RAN + Context	12	5,682	3,777	0.89	0.10

B: baseline predictor including UE category and the cell frequency band as input variables
R^2: coefficient of determination (between 0 and 1 - the closer to 1, the better)
$\tilde{E}$: median error ratio

Figure 5.3-2: Comparison of training data for LTE throughput predictions [14]

5.3.2 PF Deployed in MNO Domain

For operational aspects, such as network monitoring and customer experience management, MNOs operate several systems and analyse negative impacts on the mobile network. Such information is derived from cell-specific performance metrics, subscriber signalling information and events in the RAN and Core Network, dedicated network measurement protocols such as MDT, as well as alarming and monitoring systems and probes. Infrastructure information (antenna location, physical and software configuration) combined with geo-graphical information such as buildings, bridges and tunnels are used for PF-distribution simulations.

A combination of all mentioned data sources with real-time processing can not only create highly accurate coverage maps (MDT, RAN trace with cell triangulation, simulations), but also time sensitive capacity and latency predictions (PM counter, mobility events).

Such an option is reserved for each MNO that operates the network and requires specific integrations into all mentioned systems.

The NESQO TR concludes [5] in section 9.1: “Network Slicing in combination with the introduction of In-advance QoS Notifications and edge computing is expected to provide a better architecture support for the automotive requirements and use cases.” Such ability to combine both functions is limited to the MNO.

Furthermore, in addition to easier access to RAN and Core Network information, MNOs could deploy several endpoints (measurement points) to monitor specific paths in the network, helping with troubleshooting and collecting real network metrics used for QoS prediction because Machine Learning and AI models need training data.

Active/Passive/Hybrid Testing is a common approach to measure different network metrics like throughput, latency and jitter between two IP nodes (see for example [16]). Those nodes or measurement points could be either at the E2E side (UE and V2X AS) or some nodes in the path. An

MNO could efficiently coordinate testing sessions in order to measure network metrics at different points without disturbing all users sharing the same network. Indeed, even though traditional speeds tests, using a file download/upload, are a simple way to get an E2E achievable throughput, it is not a scalable solution. They don't allow multiple users to share the same network at the same time to train the AI models (i.e. every user performing a speed test at the same time will share their bandwidth, thus leading to false results as the measurement will affect everyone else).

Considering the current 3GPP solution discussion in 3GPP Rel. 16 [17] [18] the location of the PF within the MNO is for further study.

5.3.3 PF Location Comparison OTT/MNO

Limitations on the prediction availability and accuracy of an OTT-deployed PF are expected based on the following assumptions:

- MNO operational data mentioned in section 5.3.2 will not be shared with external parties.
- The data basis for predictions by the OTT-solution is limited to client data retrieved by high[5] and low-level APIs of the UE modem.
- The UE 'probe' penetration-rate is limited to a subset of 'vehicle-mounted UEs'.

QoS management functions are controlled and operated by MNOs. Monitoring functions to operate them, collect necessary information and provide an extensive view of the network performance are used, as outlined in section 5.3.2. It is in the interest of each MNO to have a detailed overview of its network performance. Therefore, the information depth is significantly higher than what an OTT-Player can achieve with an OTT UE-based crowd-sourcing of network information.

In summary, a PF deployed in the MNO domain is expected to provide accurate QoS predictions thanks to access to detailed network data and analytics, whereas a PF deployed in the OTT domain may face challenges in providing accurate QoS predictions due to the lack of access to relevant network data.

The big challenge for MNO's compared to an OTT PF deployment is the interoperability and general availability of a PF in each MNO network. Such a challenge could be overcome via 3GPP standardisation.

If only basic predictions on network coverage or on statistical/historical probe data are required without live network insights, an OTT deployed solution can provide – on a global scale and with a fast time-to-market – an alternative that does not require MNO interoperability and deployment of the prediction function in each network.

Table 5.3-2: Comparison between OEM and MNO deployment

[5] Examples for high-level APIs are Android [19] and iOS [20] Developer APIs.

	OTT deployment	MNO deployment
Advantage	• No geo-graphical limitation beyond what the road/vehicle network allows	• Data availability with deep network insights • UE measurements benefit from MNO wide UE footprint • Combination with network slices, QoS management and network analytics function
Limitations	• Limited to UE measurements, lacking information on network infrastructure KPIs • Partial network view due to limited number of clients compared to MNO UE footprint	• Limited to MNO footprint[5]
Challenges	• Sufficient market penetration	• Interoperability[6] • Longer time-to-market

Our observation is that, based on [5] prediction requirements for Tele-Operated-Driving, an OTT-deployed solution is **not a viable technical option**.

5.3.4 Business Environment Aspects of OTT-Based Solutions

In section 5.3.3, the pros and cons of OTT/Third-Party and MNO solutions are compared on a technical level. Several additional analyses could also be performed factoring in the business environment and perspective.

Predictions by OTT-based solutions might be implemented in several ways. One is when a certain OEM implements its own prediction platform, and another is when such a platform is offered by a Third-Party provider which then offers prediction to multiple OEMs. In the first case, the availability of prediction depends on the OEM's efforts, and its accuracy in specific areas depends on its market penetration (e.g. a few vehicles of a certain OEM in a certain area might not generate enough data/measurements to guarantee adequate accuracy in that specific area). In the case of predictions offered by a Third-Party provider to several OEMs, a point to be further discussed concerns the ownership of measurements.

Can measurements of a specific OEM be used by the Third-Party provider to generate prediction for another OEM? Even if data is anonymised (measurements do not carry information regarding, for example, original OEM), some information could still be related to specific OEM's solutions (specific UE type or UE capabilities adaptation, etc.), which might either be sensitive for the OEM or not applicable to other OEMs. This requires further investigation, bearing in mind the implications of possible agreements between a Third-Party provider and OEMs on prediction availability and associated accuracy. In fact, this might either limit the prediction accuracy offered by a Third-Party provider, or it

[6] Challenge could be overcome with a 3GPP standardised solution.

might require the Third-Party Provider in any case to collect OEM-specific measurements to support OEM-specific predictions.

5.4 Delivering QoS Prediction Notifications to V2X Applications

In order to deliver QoS Prediction notifications from the network to all V2X UEs, multiple options may arise depending on implementation choices from OEMs, MNOs or Third Parties. As already mentioned in section 5.3.3, one of the biggest challenges to increase the QoS prediction feature availability would be the interoperability between all different stakeholders. Such a challenge could be overcome via 3GPP standardisation, hiding the network complexity from the users thanks to a standardised interface. Consequently, QoS prediction notifications could be efficiently delivered to all V2X applications no matter which vendor solution is used.

NESQO TR [5] already identified a number of potential options that could be used to deliver QoS prediction notifications to consumers. Later, 3GPP SA2 and 3GPP SA6 defined new mechanisms that eNESQO has identified as potential candidates for the delivery of the QoS prediction notifications to the V2X application and V2X application server. Those mechanisms could complement the Rel-16 solution already described in [10] and [18] for the delivery of QoS prediction to the V2X application server which is based on the QoS Sustainability Analytics. While eNESQO has not fully studied how those mechanisms could be used for such a purpose, eNESQO concluded that it would be beneficial that such research is performed by 3GPP SA2 and 3GPP SA6 in order to evaluate if they can be extended or adapted for the purpose of delivering QoS prediction notifications. Specifically:

a) The mechanism defined by 3GPP SA2 based on the **extended NG-RAN Notification,** which could be used for the delivery of the QoS prediction notification to the V2X application. This mechanism is specified by [27] in Section 5.7.2.4 and [18] Section 5.4.5.3. The mechanism uses NAS signalling to inform the UE about potential changes in the QoS parameters (i.e., 5QI, GFBR, MFBR) that the NG-RAN is currently fulfilling for the QoS Flow. In the current mechanism the notification to the UE is triggered by a QoS change that has already happened, but the mechanism could be modified to be triggered by a prediction of a potential QoS change.
b) The mechanism based on the **V2X Application Enabler (VAE) layer,** which could be used for the delivery of the QoS prediction notification to the V2X application and to the V2X application server.

More information about option a) is provided in Section 5.6. and an outline of the method for option b) is described in the remaining part of this Section. Indeed, 3GPP SA6 already started to define the VAE layer to ensure optimal use and deployment of V2X Applications on 3GPP networks. The VAE capabilities should be offered as APIs to the V2X Applications. The VAE layer considers interactions on both server and client side, with the VAE server interacting with V2X application-specific server and with VAE client interacting with V2X application-specific client. Finally, server-client interaction can be offered through a VAE server-client interface, where the VAE server offers APIs allowing the VAE client to access the functionalities provided by the VAE server.

In TS 23.286 [21], 3GPP defined the functional architecture, procedures and information flows for the VAE layer over 4GS. Procedures defined for the interaction between the VAE layer and V2X applications can also be found, together with information on how the VAE layer interacts with 4G network entities. See Annex B for more details.

The ongoing work in 3GPP to enhance the VAE layer for 5GS is captured in TR 23.764 [22]. The aim of this work is to define the interfaces used by the VAE server to interact with 5G network functions, and

also to identify what enhancements are needed for the VAE layer to support the architectural and procedural improvements for V2X services defined in TS 23.287 [18].

One important topic in eNESQO concerns the delivery of QoS prediction notifications to V2X applications. The VAE layer could be exploited to this aim and currently the work in TR 23.764 [22] is addressing the issue of how to support the application layer to provide/adapt the service operation dynamically by exposing QoS information/analytics by 5GS (e.g. NWDAF). Of particular relevance for 5GAA eNESQO is the notification on Potential QoS change procedures defined for V2X services (see 3GPP TS 23.287 [18]), based on the notification on QoS Sustainability provided by NWDAF in 3GPP TS 23.288 [10]. This procedure can be considered as a baseline of QoS prediction, and the 3GPP SA6 is working on enhancing the VAE layer capabilities to support this procedure, and to expose the notification on QoS Sustainability to the V2X application layer.

5.4.1 QoS Prediction Notification from VAE Server to V2X Application Server

In TR 23.764 [22], 3GPP SA6 is currently designing a procedure for monitoring and control of QoS for eV2X communications targeting the exposure of QoS Sustainability by NWDAF to V2X application layer. The high-level flow of notifications on QoS Sustainability analytics supported by the VAE layer via the procedure for monitoring and control of QoS for eV2X communications is shown in Figure 5.4-1. In this procedure, the VAE server acts as an Application Function (AF) towards the 5GS, and in detail towards NWDAF/NEF for this particular procedure (NWDAF in case of trusted AF, NEF in case of untrusted AF, respectively). The VAE server subscribes to the QoS Sustainability analytics service from NWDAF and, if available, receives the notification from the NWDAF. After processing such information, the VAE server may provide the QoS monitoring information to the V2X application-specific server, which is then able to decide whether to trigger an adaptation based on the received notification (i.e. reaction to the reception of QoS prediction notification).

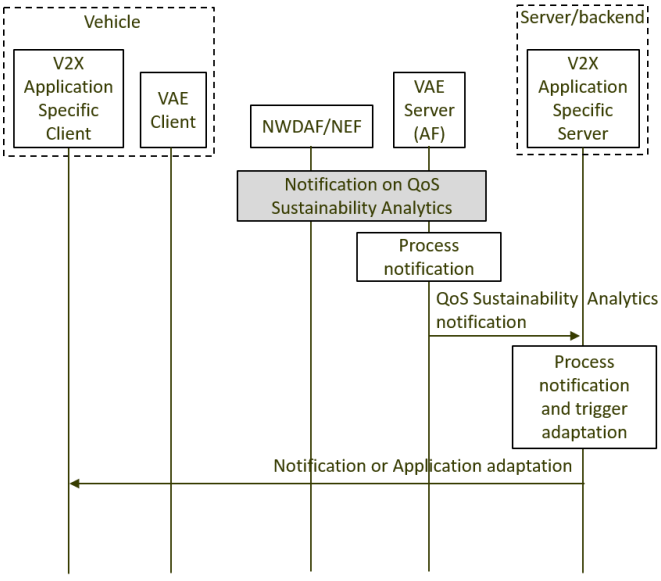


Figure 5.4-1: High-level flow of notification on QoS Sustainability analytics supported by the VAE layer via the procedure for monitoring and control of QoS for eV2X communications

The procedure for monitoring and control of QoS for eV2X communications currently being defined by 3GPP SA6 can be considered as a baseline for a standardised delivery of QoS prediction notifications to V2X applications. Please note that with the current approach in [22], the following is possible:

- *Server-based reaction to QoS prediction*. QoS prediction notification is delivered by VAE server to V2X application-specific server, which then triggers the appropriate reaction. The reaction is finally propagated towards the vehicle side, i.e. the server communicates the selected reaction to the V2X application-specific client which ultimately enforces it.
- *Server-based delivery to vehicle*. QoS prediction notification is delivered by VAE server to V2X application-specific server, which then provides the notification to V2X application-specific client. In this case, the reaction is triggered on the vehicle side once the QoS prediction notification is received from the V2X application-specific server.

As discussed above, the current procedure defined by 3GPP SA6 can enable server-based reactions to QoS prediction and server-based delivery to the vehicle. In both cases, the notification is exposed and handled firstly at the server side. This could potentially become a bottleneck for server-based delivery to vehicles, as it adds another step to delivery towards the vehicle which adds latency in the whole process of delivery. For instance, the interface between the VAE server and the V2X application-specific server may go through public Internet, depending on where the V2X application-specific server is located. This implies that the path of notification delivery may leave the MNO-controlled network. In this case when the notification leaves MNO-controlled network, going through the V2X application-specific server may be inefficient if the notification is required to be delivered towards the vehicle side with a low latency constraint.

5.4.2 QoS Prediction Notification from VAE Server to VAE Client

To overcome the issue discussed above, further enhancements could be considered for the procedure for monitoring and control of QoS for eV2X communications. Taking into consideration the availability of the V1-AE reference point between the VAE server and the VAE client, the V1-AE could be used to expose notifications on QoS Sustainability analytics to the VAE client, i.e. on the vehicle side. A high-level flow of notification on QoS Sustainability analytics to the vehicle side supported by the VAE layer via an enhanced procedure for monitoring and control of QoS for eV2X communications is depicted in Figure 5.4-2. The enhancement considers the following:

- The V2X application layer indicates that notification on QoS Sustainability analytics should be delivered to the V2X application-specific client. This triggers the VAE client to subscribe to the VAE server for reception of notifications on QoS Sustainability analytics. This could happen in a registration phase during which the V2X application layer indicates to the VAE layer whether the notification should be provided to the application layer on the server or client side. This would allow the V2X application layer, which has service knowledge, to instruct the delivery of notifications which will be executed by the VAE layer towards the server/client side based on service needs.
- Upon reception of a notification on QoS Sustainability analytics from 5GS, the VAE server processes the notification and checks whether, according to previous configurations from the application layer, the notification should be exposed to the V2X application-specific server or V2X application-specific client.
- If the notification should be exposed to the V2X application-specific client, the VAE server uses the VI-AE interface to deliver the notification on QoS Sustainability analytics to the relevant VAE client.
- Upon reception of notifications on QoS Sustainability analytics from the VAE server, the VAE client processes the notifications and exposes them to the relevant V2X application-specific client.

This enhancement allows to expose QoS Sustainability analytics towards the vehicle without the need of having the notification handled by the V2X application-specific server side, thus cutting additional delays for notification delivery. This enhancement considers as a baseline that the information of QoS Sustainability analytics is defined by 3GPP to be exposed to AF (i.e. VAE server in the VAE layer). Thus, the enhancement does not modify the already defined procedure in SA2. The enhancement exploits a user-plane-based interface (V1-AE) defined between the VAE server and the VAE client to expose the notification to the vehicle side. Thus, the considered enhancement does not require the standardisation of new interfaces or messages. It does, however, require further work in 3GPP SA6 to introduce a relevant API at the VAE server allowing the VAE client to retrieve notifications on QoS Sustainability analytics from the VAE server, together with a relevant extension of the procedure for network monitoring notifications.

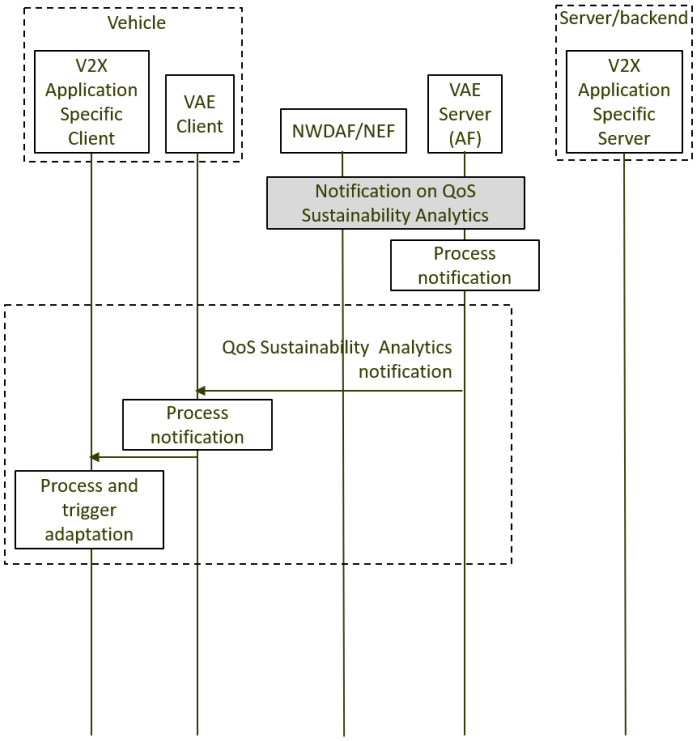


Figure 5.4-2: High-level flow of notification on QoS Sustainability analytics to the vehicle side supported by the VAE layer via an enhanced procedure for monitoring and control of QoS for eV2X communications

5.5 Application and Network Reaction to QoS Prediction

Analysis of application reactions to QoS prediction is necessary to properly understand how OEMs can take advantage of QoS prediction as well as to validate the in-advance time implications of prediction and the relevance of QoS prediction outputs for application adaptation. Moreover, before being delivered to the application, or in parallel with such procedures, prediction notification may also be used by the network itself to consider a suitable network reaction.

This section contains an analysis of application and network reactions at the reception of QoS prediction of coverage change, as well as a discussion of application adaptation in relation to the 3GPP framework.

5.5.1 Reaction to QoS Prediction of Coverage Change

In this Section, we discuss application and network reactions at the reception of QoS prediction of coverage change. In particular, we consider a prediction message containing the following information:

- The UE is predicted to leave the connectivity by means of Normal Coverage (e.g. the UE is currently served e.g. by a legacy LTE cell), with potential information about the time when the UE is predicted to leave the Normal Coverage together with information on for how long the UE is expected to not be reachable via Normal Coverage.
- The UE is predicted to be reachable by using Coverage Enhancement (CE) mode operations (e.g. the UE is predicted to be reachable via LTE-M by enabling CE-Mode A or CE-Mode B) after leaving the Normal Coverage, potentially with information about for how long the UE is expected to be reachable in CE-Mode.

We assume that the UE modem is a non-CatM UE supporting transmission by means of both Normal Coverage capability and CE-Mode capability. It is also assumed that the network is able to perform UE-specific authorisation for the utilisation of CE-Mode. To this aim, several approaches could be used:

- Utilising the Subscriber Profile ID (SPID) associated with the UE.
    - One option (suggested in [23]) is to re-use the SPID value 253 'Automotive device subscriber' whose primary purpose is to identify a device that is permitted to use two Rx antenna ports for NR instead of four. A side-effect of this option is that all UEs associated with SPID value 253 will be considered as UEs authorised for switching the utilization of CE mode (i.e. it would not be possible to distinguish a UE which is 'only' a UE that is permitted to use two Rx antenna ports for NR but it is not requiring or not permitted to use the switching of CE mode utilization).
    - One option is to use an operator-specific SPID value to identify the UEs which are both 'Automotive device subscriber' and authorised for switching the utilisation of CE-Mode (and then using the SPID value 253 for identifying the UEs that are 'Automotive device subscriber' but are not allowed or do not require the switching of CE mode utilization). A side-effect of this option is that the utilisation of operator-specific SPID values might have issues in roaming cases (i.e. different operators associating different meaning to a specific SPID value), thus requiring further configuration and/or agreements among operators for roaming.
- Utilising the Enhanced Coverage Restricted parameter [24].
    - 3GPP included the Enhanced Coverage Restricted parameter as a part of UEs' subscription data in the Home Subscriber Server (HSS), which can be then used by operators to enable/disable specific subscribers from Enhanced Coverage features. This option requires support in the core network of the Enhanced Coverage Restricted parameter and associated procedures.

In order to understand how applications and networks might react to the reception of a prediction of coverage change, it is important to discuss how application and network sides behave differently according to the coverage capabilities. Different states can be associated at application and network sides. Examples of such states can be:

- **States at application side** (both on the vehicle and server side).
    - *Normal Coverage.* This state reflects that the vehicle can take advantage of normal connectivity, then all applications/services are enabled, including applications ranging from small data transfer of delay-tolerant information to applications requiring, for instance, high bit rates, low latency, large amount of data transfer, etc. Examples of such applications are: basic sensor reporting from the vehicle, massive sensor reporting from the vehicle, video streaming from the vehicle, remote control, basic connectivity

for status checks from the remote server, software updates from remote servers, HD map acquisition from remote servers, infotainment services from remote servers, etc.
    - *Coverage Enhancement.* This state reflects that the vehicle can exploit only limited connectivity by using CE mode, then only a reduced set of applications/services are enabled. For example, applications associated with small data transfer of delay-tolerant information while applications with higher QoS requirements are disabled. Examples of applications allowed in this configuration are: basic sensor reporting from the vehicle and basic connectivity for status checks from remote servers.

NOTE: in the examples above, the definition of states on the application side are associated with connectivity requirements of the application. States on the application side could also be defined according to a vehicle's features, e.g. a state could be a 'parked vehicle' requiring basic sensor reporting and another one could be a 'moving vehicle' requiring all applications/services. Additional vehicle-related states could be considered as well and then applied to application states (e.g. a 'moving vehicle' state could be associated either with a *Normal Coverage* application state or to *Coverage Enhancement* application state according to connectivity capabilities. Additional examples of states on the application side as well as additional states at vehicles side are left FFS.

- **States at network side.**
    - *Normal Coverage*. In this case, the UE operates in a normal mode and the network does not enforce any particular behaviour on the UE (in addition to the usual configuration for admission control, QoS management, priority, etc.).
    - *Coverage Enhancement*. The network enforces CE-specific features which are in line with the UE operating in CE-Mode in order to, for example, limit the amount of transmitted data, reduce the maximum allowed bit rate of transmission, and so on. Examples might include specific policies for rate limitation, different charging, changes in UE/traffic priority, and other changes of UE-related information for traffic treatment to comply with operations in CE-Mode.

Considering the list above, there could be changes from *Normal Coverage* state to *Coverage Enhancement* state (or vice versa), triggered from the application or network side. From this perspective, QoS prediction can be used by applications or networks to understand in advance to which state to switch to and to timely adapt its behaviour to cope with the new target state. One can also note that the selection of a certain state on the application side could also impact the state at network side (and vice versa). For instance, a *coverage enhancement* state on the network side should involve the setting of *coverage enhancement* state at application side too, to avoid a misbehaviour at the application side (e.g. sending huge amounts of data that might be charged in a different way when transmitted via CE-Mode operations or trying to access some services/applications which might not be available when a UE is reachable via CE-Mode operations).

The switching between the states listed above is driven by some **triggering conditions** including for example:

- **Application-based triggering conditions.**
    - State change from *Normal Coverage* to *Coverage Enhancement*: (i) the vehicle is parked (or expected to be parked within a short time); (ii) at the target location *Normal Coverage* is not available (or predicted to be not available) but connectivity via CE is available (or predicted to be available); (iii) the vehicle requires basic sensor reporting services or basic connectivity for status checks from remote servers. When these three conditions are met, a state switch from *Normal Coverage* to *Coverage Enhancement* is triggered.
    - State change from *Coverage Enhancement* to *Normal Coverage*: (i) the vehicle is moving (or predicted to be moving); (ii) the vehicle can be (or predicted to be) reached via

Normal Coverage in the area where the vehicle is (or will be) moving; (iii) the vehicle requires other applications in addition to basic sensor reporting services or basic connectivity for status checks from remote servers. When these three conditions are met, a state switch from *Coverage Enhancement* to *Normal Coverage* is triggered.

- **Network-based triggering conditions.**
  - State change from N*ormal Coverage* to *Coverage Enhancement*: (i) it is predicted that the UE will not be reachable by means of *Normal Coverage* and that the unreachability is expected to last longer that a certain time interval (for instance, if the network wants to prevent switching to operations in CE-Mode in case of a temporary loss of *Normal Coverage* of, for example, only one second); (ii) the UE is predicted to be reachable by using CE-Mode; (iii) the network is expected to have enough resources to serve the UE using CE-Mode. When these three conditions are met, a state switch from *Normal Coverage* to *Coverage Enhancement* is triggered.
  - State change from *Coverage Enhancement* to *Normal Coverage*: (i) the UE is moving in a location where *Normal Coverage* is available (or predicted to be available); (ii) the network is expected to have enough resources to serve the UE using *Normal Coverage* operations. When these two conditions are met, a state switch from *Coverage Enhancement* to *Normal Coverage* is triggered.

NOTE: state configurations and triggering conditions are **not limited to the examples above**. For instance, additional state triggering conditions might be defined to handle coverage mode switching for a moving vehicle which, in the event of a loss of normal connectivity, may switch to *Coverage Enhancement* configuration with either application- or network-driven reactions, to maintain basic sensor reporting services while moving. Additional examples and descriptions of state configurations and triggering conditions are FFS.

In addition to state configurations on the application and network side and triggering conditions for state adaptation, it is important to analyse the **reactions** generated on both application and network sides at the reception of a QoS prediction message, reactions that will then allow the application or the network to switch from one state to another. An example of state switching for application-based reaction is depicted in Figure 5.5-1, while an example of state switching for network-based reaction is depicted in Figure 5.5-2.

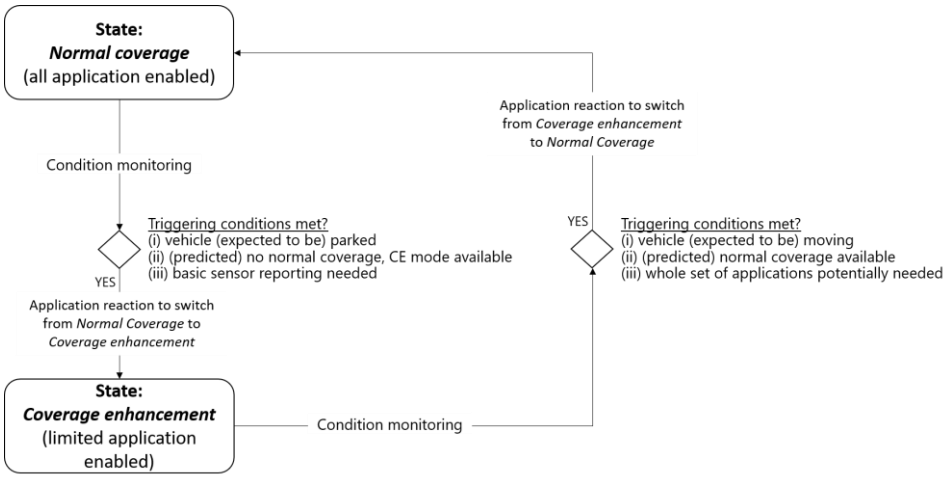


Figure 5.5-1: State switching for application-based reaction

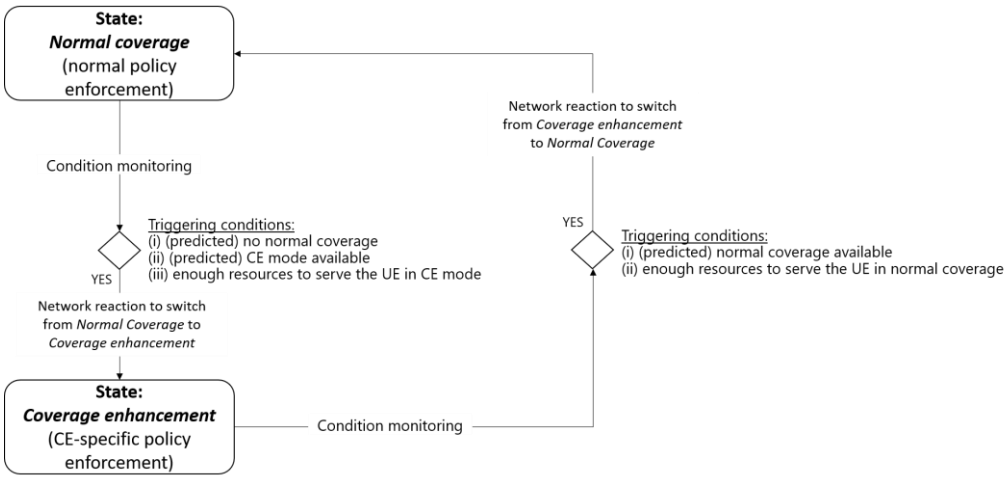


Figure 5.5-2: State switching for network-based reaction.

In the following, more details on application- and network-based reactions are provided. It is considered the case when it is predicted a coverage change from *Normal Coverage* to *Coverage Enhancement* (although the analysis can be extended also for predictions of changes from *Coverage Enhancement* to *Normal Coverage* by properly adapting state configurations, triggering conditions, and reactions). In Section 5.5.1.1 it is considered that the application at the vehicle side is the receiver of the prediction and that the application triggers the switch of coverage change, while in Section 5.5.1.2 it is considered that the network is the receiver of the prediction which then triggers the switch of coverage change.

In the following, it is considered that the *Normal Coverage* state on the application side is mapped to the fact that the UE modem on the vehicle side utilises *Normal Coverage* capabilities, while the *Coverage Enhancement* state on the application side is mapped to the fact that the UE modem on the vehicle side utilises the CE-Mode capability (and this is of course reflected on the network side).

In the examples in the remainder of this section we consider:

- The Prediction Function (PF)
- Vehicle-side, comprising the application(s) and the UE modem
- Backend-side, marked as OEM cloud
- Mobile network, comprising both RAN and core sides

5.5.1.1 Application-Driven Reaction to Prediction of Coverage Change

In this case, it is assumed that the recipient of the prediction is the application side (in the considered example, the application at vehicle side). The reception of a prediction of coverage change by the application on the vehicle side generates a reaction involving a state change at UE and OEM's cloud sides, which also generates a state change at network side.

The following assumptions are considered in the example presented in this Section. The UE modem indicates to the network its preferred transmission mode (i.e. whether *Normal Coverage* or CE-Mode operations should be used) by modifying the UE capabilities [25] signalled during the attach procedure [24] to the mobile network. This implies that the UE modem should support configurations from higher layers to update the configuration of the utilization of CE-mode capability.

For both application and network sides, the initial state is *Normal Coverage* (at UE modem side, this is reflected by the fact that the CE-mode capability is not set). The flow diagram of application reaction to the reception of QoS prediction of coverage change is depicted in Figure 5.5-3.

- The application on the vehicle side receives the QoS prediction of coverage change from the PF, indicating that the vehicle is predicted to lose *Normal Coverage* within a certain time interval of X seconds and that coverage through *Coverage Enhancement*s operations is predicted to be available.
- The application on the vehicle side checks whether the triggering conditions to adapt its state are met. From the example in the Section above, we have three conditions to be met: (i) the vehicle is parked (or expected to be parked within a short time); (ii) the vehicle is expected to be in a location where *Normal Coverage* is not available while connectivity using CE-Mode is expected to be available; (iii) the vehicle requires basic sensor reporting services when parked. The reception of QoS prediction of coverage change from *Normal Coverage* to *Coverage Enhancement* helps to check point (ii) of the triggering conditions. The vehicle-side application then checks points (i) and (iii); if they apply the application decides to trigger its reaction to switch its state from *Normal Coverage* to *Coverage Enhancement*.

NOTE: the information included in the prediction (e.g., when it is predicted to lose *Normal Coverage*) can be used by the application on the vehicle side in several ways; for example, to timely notify the backend-side about the upcoming loss of *Normal Coverage* (in case the backend-side was not another receiver of the prediction), to safely (before the switch of connectivity mode) modify the vehicle's behaviour (application configurations, sensor reporting, warning/messages to drivers, etc.), and to decide the most adequate time to trigger the reaction (e.g. after successful adaptation). Potentially, the application might decide to trigger the reaction after the loss of *Normal Coverage* is detected, and to exploit the prediction information to have enough time to safely adapt the vehicle's behaviour before losing *Normal Coverage* connectivity.

- The application on the vehicle side triggers the reaction to change the state from *Normal Coverage* to *Coverage Enhancement*. In this example, we consider that the reaction is that the vehicle-side application initiates a procedure to change the chipset capability signalled from the UE modem to the network. If the UE is currently attached, the signalling of a change of UE capability would require the UE to firstly detach from the network and then perform a new attachment. As a consequence, in this example the reaction can be described as follows: (i) the UE modem performs a detachment from the mobile network upon request from the application; (ii) the application configures the UE modem to set the CE-mode capability as supported; (iii) the UE modem triggers an attachment to the mobile network upon request from the application, where the new capability (CE-Mode supported) will be signalled. The attachment triggers the network to change its state from *Normal Coverage* to *Coverage Enhancement.*
- When the attachment is completed, the vehicle-side application completes its transition to *Coverage Enhancement* state. It is then configured to use only applications associated with basic sensor reporting.
- Once the vehicle-side state is updated, the application at vehicle-side informs the other communication end-point(s), e.g., its own OEM's cloud, about the state currently in use. This allows the OEM's cloud to adapt its state as well, (e.g. disable all applications except basic connectivity for the vehicle's state check).

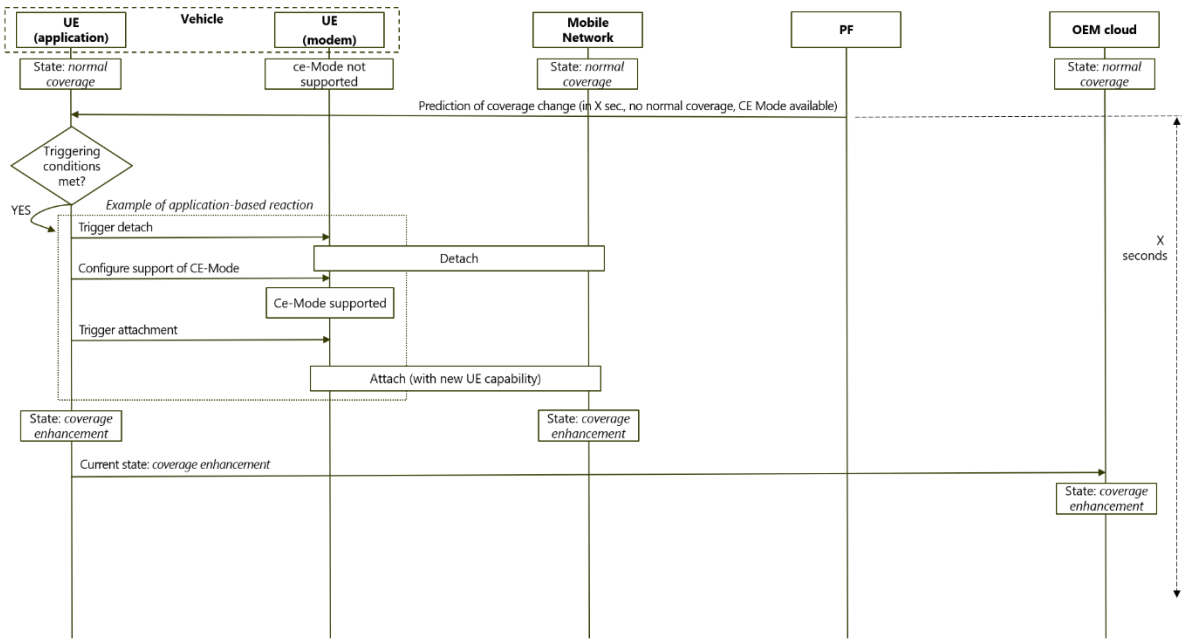


Figure 5.5-3: Application-driven reaction to QoS prediction of coverage change (reaction to change state from *Normal Coverage* to *Coverage Enhancement*)

5.5.1.2 Network-Driven Reaction to Prediction of Coverage Change

In this case, it is assumed that the recipient of the prediction is on the network side. The reception of a prediction of coverage change by the network generates a reaction involving a state change at network side, which also generates a state change on the application side (vehicle and OEM's cloud).

The following assumptions are considered in the example presented in this Section. The network is already aware that the UE is able to support operations in CE-Mode (i.e. the UE already signalled its capability of supporting operations in CE-Mode) and that the network is able to configure the utilisation of CE-Mode. This could be achieved using RRCConnectionReconfiguration procedure [26] where the CE-Mode element (included in the PhysicalConfigDedicated element carried by the radioResourceConfigDedicated element) can be set to 'setup' (e.g. 'CE-Mode A') or to 'release'. In the former case, the network triggers a physical channel re-configuration configuring the UE to use, for example, CE-Mode A. In the latter case, the network triggers a physical channel re-configuration configuring the UE to release the utilisation of CE-Mode.

For both application and network sides, the initial state is *Normal Coverage*. The flow diagram of network reaction to the reception of QoS prediction of coverage change is depicted in Figure 5.5-4.

- The network receives the QoS prediction of a coverage change from the PF, indicating that the UE is predicted to lose *Normal Coverage* within a certain time interval of X seconds and that coverage through *Coverage Enhancement*s operations is predicted to be available.
- The network checks whether the triggering conditions to adapt its state are met. From the description in the Section above, we have three conditions to be met: (i) it is predicted that the UE will not be reachable by means of *Normal Coverage* and that the unreachability is expected to last longer that a certain time interval; (ii) the UE is predicted to be reachable by using CE-Mode; (iii) the network is expected to have enough resources to serve the UE using CE-Mode. The reception of QoS prediction of coverage change from *Normal Coverage* to *Coverage Enhancement* helps to check point (i) and (ii) of the triggering conditions. The network then

checks point (iii), and if this point also applies then the network decides to trigger its reaction to switch its state from *Normal Coverage* to *Coverage Enhancement*.

NOTE: the information included in the prediction (e.g., when it is predicted to lose *Normal Coverage*) can be used by the network in several ways; for example to decide the most adequate time to trigger the reaction (e.g. after successful adaptation of its behaviour to successfully accommodate the upcoming UE operating in CE-Mode). Another possible configuration is also that the application is a receiver of the prediction, such information can be used by the application to safely modify its behaviour before the loss of normal connectivity, for instance adapting application configurations and sensor reporting, displaying warning/messages to drivers, etc.

- The network triggers the reaction to change the state from *Normal Coverage* to *Coverage Enhancement*. The reaction considered in this example is that the network initiates the RRCConnectionReconfiguration procedure marking the CE-Mode element (included in the PhysicalConfigDedicated element carried by the radioResourceConfigDedicated element) as 'setup' (e.g. to CE-Mode A). This procedure triggers the network to change its state from *Normal Coverage* to *Coverage Enhancement.*
- The completion of the reaction triggered by the network also involves a state adaptation at the application side. In this case, two options can be considered.
    - Option 1. The information about the physical channel re-configuration with the enabling of utilisation of CE-Mode, is exposed on the vehicle-side application. In this case, the UE modem is able to monitor the change of CE-Mode state and *exposes* such information to the vehicle's application, which adapts its state to *Coverage Enhancement*. Once the vehicle's state is updated, the vehicle-side application informs the other communication end-point(s), such as its own OEM's cloud, about the state currently in use. This allows the OEM's cloud to adapt its state as well.
    - Option 2. The network exposes the information about the change of CE-Mode utilisation (i.e. that the UE is now operating in CE-Mode), e.g., via the NEF towards the OEM's AF. At the reception of such information, the OEM's AF could further inform OEM's cloud about this change and OEM's cloud could trigger the update of state to *Coverage Enhancement*. Once the OEM's cloud state is updated, the OEM's cloud could inform the application at vehicle-side about the state currently in use, vehicle which then adapts its state to *Coverage Enhancement*.

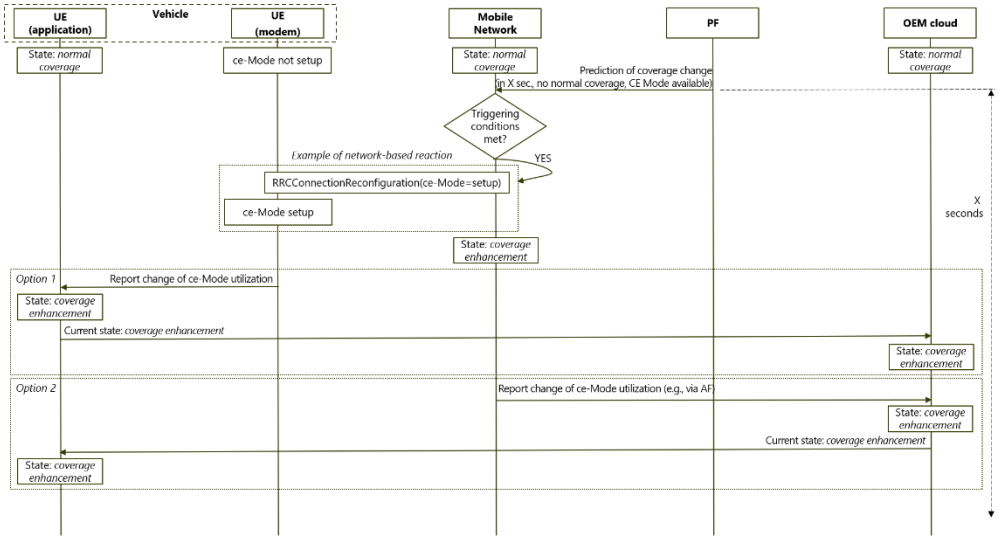


Figure 5.5-4: Network-driven reaction to QoS prediction of coverage change (reaction to change state from *Normal Coverage* to *Coverage Enhancement*)

5.5.2 Application Adaptation and 3GPP QoS Framework

Application design should consider the capability of adapting the application's behaviour (e.g. encoding, flow priority, packet inter-arrival time) in order to withstand changes in experienced network performance across time and space while a vehicle is moving, and changes due to variations of network load, radio link quality, etc.

The adaptation can be also supported from the network by enforcing specific QoS treatments; for example, considering requirements on the lowest acceptable and highest needed bit rates, among the parameters defined by 3GPP in [27] associated with specific QoS Flow treatment. Taking the bit rate example, 3GPP defines for 5G systems the Guaranteed Flow Bit Rate (GFBR, the bit rate that is guaranteed to be provided by the network to a Guaranteed Bit Rate (GBR) and Delay Critical GBR QoS Flow, over the Averaging Time Window) and the Maximum Flow Bit Rate (MFBR, limiting the bit rate that is expected by a GBR QoS Flow). Considering these two values, the network guarantees a bit rate up to the GFBR, it tries to fulfil bit rates from GFBR up to MFBR (considering that the traffic above GFBR may be delayed more than the maximum latency constraint), and the traffic exceeding the MFBR may get discarded or delayed by a rate shaping or policing function at the UE, RAN, UPF. As a consequence, GFBR and MFBR can be mapped according to the requirements on the lowest acceptable and highest needed bit rates of the application. A similar behaviour can be also found in 4G systems, where the GBR and the Maximum Bit Rate (MBR) parameters of the 4G QoS model can be mapped, respectively, to GFBR and MFBR of the 5G QoS model.

The guidelines of QoS framework utilisation for application adaptation is, for instance, considered in 3GPP for Framework for Live Uplink Streaming (FLUS) [28], which is assumed to be an adaptive application where the source can adjust the transmission bit rate to the currently measured/estimated link bit rates. This can be achieved by influencing the encoder bit rate or by dropping frames before transmission, for instance through the utilisation of rate adaptation algorithms such as Self-Clocked Rate Adaptation for Multimedia (SCReAM) [29] [30]. The expectation is that the system delivers, as often as possible, a certain target quality, and that a quality higher than the target is not needed. Depending on the video codec configuration (Codec Profile, Codec Level and Encoder Features), the video quality is associated with a bit rate of the compressed stream. In an example in [28], here depicted in Figure 5.5-55.5-5 (left-hand side), a resulting video bit rate of ~15Mbps corresponds to the target video quality ('as expected quality'). A resulting bit rate above 4Mbps corresponds to an 'ok' quality (quality is not perfect, but still good to use). A resulting bit rate between 800kbps and 4Mbps is the 'better than nothing' quality, where the video quality contains very obvious quality artefacts. When the resulting bit rate is below 800kbps, the video quality is 'unusable' (i.e. effectively no video transfer).

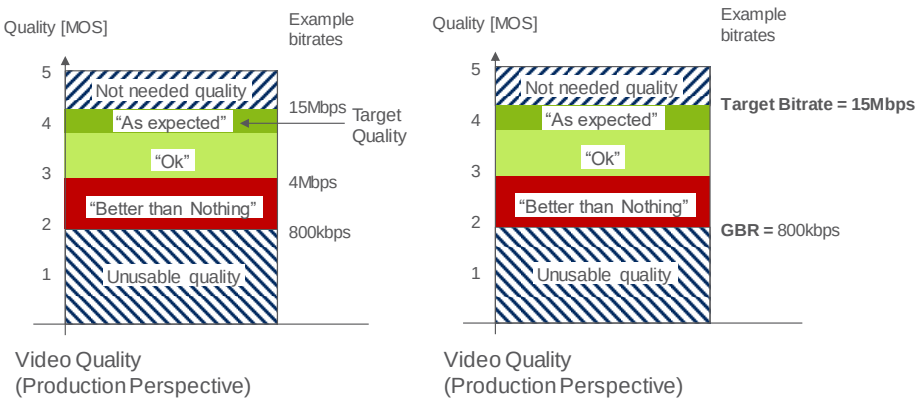


Figure 5.5-5: Quality perspective (left) and target QoS boundaries mapped to a 4G QoS model (right) for FLUS as indicated in [28]

3GPP discussed a desired QoS Flow behaviour to support a FLUS application in [28], considering the aspects associated with traffic treatment when the network enforces GFBR and MFBR. A first aspect to consider is that the admission control algorithms are going to reject/pre-empt a QoS bearer based on the GFBR value. Therefore, the utilization of high values for GFBR may increase the risk of the system admission control rejecting the QoS Flow (please note that some network events, for instance handovers to other cells/other access networks, may re-trigger the admission control process). Therefore, low values for GFBR may be preferable (and it should also be considered that high values for GFBR are associated with higher cost). A further aspect to consider is traffic treatment when the bit rate exceeds the GFBR. When a GBR Flow is admitted, scheduling algorithms shall guarantee the fulfilment of the GFBR by assigning adequate scheduling priority to fulfil the QoS target of the GFBR, but a behaviour for scheduling priority when traffic exceeds the GFBR is not defined and it is thus left to network implementation [27] [28]. In some implementations aiming at avoiding jeopardization of resources caused by GBR Flows, the network may decrease the scheduling priority of these flows when their bit rate is above the GFBR. In this case, as for instance mentioned in [28], the scheduling priority of the flow may be thus treated in a similar way as for 'best effort' traffic (or even with a lower scheduling priority) once the bit rate is above GFBR. In other implementations, scheduling priorities for GBR Flows exceeding the GFBR may be kept higher than the scheduling priority of 'best effort' traffic until the bit rate reaches (or it is close to) the MFBR. In this case, the scheduling priority for the QoS flow would gradually decrease with the increasing bit rate (under the assumptions that values of GFBR and MFBR are not too close). This would allow the system to still prioritise the QoS Flow when the bit rate is above GFBR, allowing the video streaming to reach the 'as expected quality' level. The conclusion from the study in [28] is that the desired QoS Flow behaviour for FLUS applications indicates that the GFBR value should be selected as the lowest acceptable bit rate (i.e. 800kbps), while the MFBR should be much larger than and close to the preferred service operation point (i.e. the target bit rate). This setting of target QoS boundaries considering 4G QoS model is depicted in Figure 5.5-55-5 (right-hand side).

According to the above, it could be beneficial to perform an analysis of different qualities for automotive applications and their relationship with experienced network conditions. Of course, the quality analysis should consider additional aspects compared to human-oriented applications, including impacts on driving and/or vehicle behaviour taking into consideration the current driving and/or vehicle behaviour (e.g. a bit rate degradation to 'better than nothing' quality might have less impact on slow-moving vehicles but a higher impact on those moving at higher speeds). This analysis would help to define the relationships with the 3GPP QoS framework (e.g. definition of target QoS boundaries associated with different application qualities) as well as the relationship with QoS prediction (e.g. definition of adequate in-advance notification intervals and the threshold for prediction generation vis-à-vis its impact on driving and/or vehicle behaviour and associated completion times).

5.6 3GPP Rel-16 Solution: Areas of Improvement

5GAA [5] has defined requirements and mechanisms for the network and application to exchange information on predicted QoS changes, in order to enable application adaptation in reaction when such information is received from the network. 5GAA has defined the message provided by the network containing the QoS prediction information as 'In-advance QoS Notification (IQN)'. Such a message is generated by a Prediction Function (PF) and delivered to an IQN Consumer which consumes the IQN performing or initiating the reaction/adaptation. The entity that delivers the IQN to the IQN Consumer is defined as the IQN Producer.

Following to 5GAA requirements, 3GPP has defined a solution in Rel-16 for application adjustment in case of notification on QoS Sustainability Analytics is received from the 5GS [10], [18]. This solution enables an Application Function (AF) to receive information about potential QoS changes from 5GS. The procedure for notification of QoS Sustainability Analytics to the V2X application server is defined in [18].

In such procedure the AF is a V2X application server that can initiate application adjustment when receiving a notification on QoS Sustainability Analytics from 5GS.

The following sections provide a description of the identified gaps between 5GAA-defined requirements [5] and the current 3GPP Rel-16 solution [10], [18]. Those gaps are addressed as areas of improvement for the 3GPP Rel-16 solution.

5.6.1 Scope Alignment

Before proceeding with the analysis it is wise to consider the difference in scope between 5GAA requirements and the 3GPP Rel-16 solution. 5GAA [5] has covered requirements and mechanisms for predictions at both application-level (End-to-End scope) and network-level (3GPP System scope). 5GAA has also considered that PF could be located both in the OTT and in the 3GPP network, and it has discussed advantages and limitations of both approaches, see section 5.3. Recent contributions also introduced the possibility for an additional PF in the UE to complement a PF in the 3GPP network [31]. The 3GPP Rel-16 solution is regarded as network-level prediction with the PF located in the 3GPP network. Specifically, in 3GPP Rel-16, PF functionality is covered by the Network and Data Analytics Function (NWDAF).

5.6.2 Architecture Considerations

For network-level prediction, 5GAA [5] has considered the following delivery options that determine the location of the IQN Producer according to the selected IQN Consumer (V2X application or V2X AS):

Table 5.6-1: 5GAA identified delivery principles, according to IQN Consumers and Producers

Option	IQN Consumer	IQN Producer	Reference point	Protocol	Notes	Rel-16 Support
1	V2X application (via the UE)	AMF/SMF/PCF(1)	N1	NAS	Under critical radio conditions which may impact radio resource availability for PDU Session, IQN may still be delivered even when user-plane resources have been removed from the PDU Session	Not supported
2	V2X application (via the UE)	RAN	Uu	RRC	Under critical radio conditions which may impact radio resource availability for PDU Session, IQN may still be delivered even when user-plane	Not supported

					resources have been removed from the PDU Session	
3	V2X application (via VAE layer)	AMF/SMF/PCF/ NWDAF	V1-AE, Vc	SBI/HTTP	V1-AE is the reference point between the VAE server and the VAE client. Vc is the reference point between the VAE client and the V2X application specific client	VAE layer for 5GS not available in Rel-16
4	V2X application server (V2X AS)	NEF(2)	N33 (Nnef)	SBI/HTTP	N33 is the reference point between NEF and AF (V2X AS)	Supported
5	V2X application server (via VAE layer)	AMF/SMF/PCF/ NWDAF	Vs	SBI/HTTP	Vs is the reference point between the VAE layer and the V2X application-specific server	VAE layer for 5GS not available in Rel-16.

(1) NAS is the protocol used in the 5GS reference point N1 between the UE and 5G Core Network control-plane functions. 5GAA [5] did not specify which NF should be identified as IQN Producer in the event NAS is used to deliver IQN. According to the 5G System Architecture, the NFs (AMF, SMF and PCF) are candidate options because they are involved in Session Management procedures. Later 3GPP introduced [27] an improved Notification Control procedure where, in case of unfulfilment of the QoS flow, NG-RAN includes more information in the message that is sent towards the SMF, as specified by [27] in Section 5.7.2.4 – where NAS signalling is used to inform the UE about potential changes in the QoS parameters (i.e. 5QI, GFBR, MFBR) that the NG-RAN is currently fulfilling for the QoS Flow. This mechanism is already used for V2X application adaptation, as specified in Section 5.4.5.3 of [18] 'QoS Change Based on Extended NG-RAN Notification to Support Alternative Service Requirements'. A possibility to enable potential QoS change notification to the UE is the usage of the already existing (or modified) Notification Control procedure where the notification itself is triggered by the prediction of a potential QoS change instead of a QoS change that has already happened. Further studies would be required to analyse the implications and impact on NAS signaling due to this extension of Notification Control procedure.

(2) 5GAA [5] also included the possibility of PCF delivering IQN according to N5 (Npcf). However, this option has been excluded by 3GPP and is not listed in the table above, since the prediction functionality has now been included in NWDAF.

In 3GPP Rel-16 solution, Options 1 and 2 are not supported, while Option 3 is implemented with PF located in the NWDAF and prediction information by means of QoS Sustainability Analytics being delivered by NWDAF via NEF to AF (V2X AS). In 3GPP Rel-16 Solution, 5GS cannot deliver the prediction

information directly to the vehicle (UE or V2X application). The delivery to the vehicle (V2X application) can happen via the V2X AS that is out of 3GPP's scope.

According to [5], the ultimate intended recipient of the IQN – and the entity that is supposed to react in the vehicle – is the V2X application in the UE [5]. Therefore, when the IQN Consumer is the V2X application both IQN Producer and IQN Consumer fall outside the 3GPP network, being the IQN Producer the AF (V2X AS). As a consequence, no applicable 3GPP standard interface is available for IQN delivery to the vehicle according to the current 3GPP Rel-16 solution. There is the V1 reference point between the V2X application in the UE and in the V2X application server in TS 23.287, but this reference point is out of 3GPP's scope.

In summary, according to the current 3GPP Rel-16 solution, IQN delivery from the V2X AS to the V2X application in the vehicle is out of scope of 3GPP specifications (i.e. there is no applicable 3GPP standard interface for IQN delivery to the vehicle).

It can be assumed that this approach is good enough when time horizons for the prediction are in the order of a few minutes or higher. For time horizons in the order of few seconds, other mechanisms such as the one described in Option 1 could be evaluated as they could potentially perform better under critical radio conditions.

Furthermore, as reflected in options 3 and 5, the exploitation of the VAE layer can be considered as a standardised method defined by 3GPP for IQN delivery to the V2X application layer.

In option 3, eNESQO has considered the exploitation of the VAE layer for prediction delivery from the VAE server towards the V2X application server side. In option 5, eNESQO has considered possible enhancements for the exploitation of the VAE layer to deliver predictions towards the V2X application on the client side. To this aim, options 4 and 5 could be further evaluated during Rel-17 3GPP work.

Area of improvement 1

The following area of improvement has been identified:

The 3GPP Rel-16 solution does not currently enable delivery of potential QoS change notifications to the vehicle for UE-side application adaptation. The 3GPP Rel-16 solution supports notification of potential QoS change to the V2S AS, which may share the potential QoS change notification with the UE-side of the application using user-plane which is not a 3GPP-standardised interface.

Potential improvements for future 3GPP consideration include delivery of potential QoS change notifications to the vehicle (UE-side) for application adaptation according to the following options:

- **The usage of a modified extended NG-RAN Notification, as specified by [27] in Section 5.7.2.4 and [18] Section 5.4.5.3; where NAS signalling is used to inform the UE about potential changes in the QoS parameters (i.e. 5QI, GFBR, MFBR) that the NG-RAN is currently fulfilling for the QoS Flow, where the notification to the UE is not triggered by a QoS change that has already happened but by a prediction of a potential QoS change (option 1 of Table 5.6-1).**
- **The usage of VAE layer to deliver predictions towards the V2X application client side (option 3 of Table 5.6-1). This option is discussed in detail in Section 5.4.2.**

5.6.3 Considerations on Supported KPIs

In the case of network-level prediction with the PF located in the 3GPP network, 5GAA [5] has defined requirements in terms of KPIs that should be supported for QoS prediction. Those KPIs are summarised in Table 5.6-2, together with current status of support in current 3GPP Rel-16 solution.

Table 5.6-2: KPIs required for QoS prediction by NESQO/eNESQO, corresponding measurement point in the 3GPP Rel-16 solution and support for prediction

End-to-End QoS KPI according to 5GAA requirements	3GPP Rel-16 corresponding Measurement Point(s) (QoS characteristic, QoS parameter or UP connection state)	3GPP Rel-16 Measurement Point Applicability (either QoS Flow or PDU Session)	3GPP Rel-16 QoS Flow type Applicability	3GPP Rel-16 Support for Prediction
Latency	PDB (QoS characteristic)	QoS Flow	GBR or Non-GBR QoS Flow	No(2)
Packet Delivery Ratio	PER (QoS characteristic)	QoS Flow	GBR or Non-GBR QoS Flow	No(2)
Uplink Throughput	• Minimum required uplink bit rate for a GBR QoS Flow: UL GFBR (QoS parameter) • Maximum required uplink bit rate for a GBR QoS Flow: UL MFBR (QoS parameter)	QoS Flow	GBR or Non-GBR QoS Flow	Yes, partially(1)
Downlink Throughput	• Minimum required downlink bit rate for a GBR QoS Flow: DL GFBR (QoS parameter) • Maximum required downlink bit rate for a GBR QoS Flow: DL MFBR (QoS parameter)	QoS Flow	GBR or Non-GBR QoS Flow only	Yes, partially(1)

Coverage/Capability (e.g. Connectivity Service Status)	UP connection state (active or inactive)	PDU Session	N/A	No(3)

(1) According to [10] Reporting Threshold(s) indicate conditions on the level to be reached for the reporting of the analytics (i.e. to 'discretise' the output analytics and trigger the notification). The level(s) relate to the QoS KPIs (i.e. the RAN UE Throughput, the QoS Flow Retainability, etc. for the relevant 5QI(s) defined in TS 28.554 [32]).
[10] include support for prediction only for RAN UE Throughput in the event of non-GBR QoS flows. RAN UE Throughput is defined as "Average UE bit rate in the cell (payload data volume on RLC level per elapsed time unit on the air interface, for transfers restricted by the air interface), per timeslot, per cell, per 5QI and per S-NSSAI". Prediction for End-to-End throughput, as suggested in [5] Section 6.4, is not currently supported in [10]. End-to-End has to be considered according to measurement points available in the 3GPP System, e.g. IP, PDCP or SDAP level.

(2) Although in [10] Section 6.9.1, it is possible to include as QoS requirements PDB and PER in the Analytics Request or Subscription, it is not possible to apply Reporting Threshold(s) to those KPIs. Also [10] did not identify any OAM parameter in TS 28.554 [32] as input data to provide predictions on PDB or PER.

(3) UP connection state has been indicated by NESQO as a possible QoS KPI to be used to provide coverage/service capability information. Coverage/Capability is to be intended as an indication of the availability/unavailability of the connectivity service required by the application. Potentially alternative ways could be defined in order to provide the consumer information about Coverage/Capability prediction.

According to 3GPP TS 23.501, services using Non-GBR QoS Flows should be prepared to experience congestion-related packet drops and delays. Also, as the NESQO identified use cases that are sensitive to both throughput and latency variation; it is expected that they will be mapped on GBR or Delay-Critical GBR use cases. **Therefore, GBR QoS Flow prediction has to be prioritised over non-GBR QoS Flow prediction.**

<u>Area of Improvement 2</u>

The following area of improvement has been identified:

3GPP Rel-16 solution does not specify prediction for the following End-to-End KPIs:

- **Latency for GBR or Non-GBR QoS Flows.**
- **Packet Delivery Ratio for GBR or Non-GBR QoS Flows.**
- **Uplink Throughput for GBR QoS Flow and Non-GBR QoS Flows (Prediction for the partial metric RAN UE Throughput is currently supported only for Non-GBR QoS Flows).**
- **Downlink Throughput for GBR QoS Flow and Non-GBR QoS Flows (Prediction for the partial metric RAN UE Throughput is currently supported only for Non-GBR QoS Flows).**
- **Coverage and Capability.**

It has to be noted that:

- **End-to-End has to be considered considering measurement points available in the 3GPP System.**
- **For GBR QoS Flows, as long as the GBR is guaranteed, the QoS KPIs latency, packet delivery ratio, uplink throughput and downlink throughput are also guaranteed by the network. As the current 3GPP Rel-16 solution supports the prediction of the QoS Flow Retainability KPI for GBR QoS Flows, the consumer does not need a prediction for the above-mentioned KPIs for the time intervals for which the GBR is predicted to be**

guaranteed. A prediction of the above-mentioned KPIs could be relevant for the time intervals for which it is predicted that GBR may not be guaranteed. This would enable the consumer to know more about the type of potential QoS change that may occur.

Due to the nature of V2X services, it is expected that GBR QoS Flow prediction has to be prioritised over Non-GBR QoS Flow prediction.

5.6.4 Other Aspects

5GAA [5] has described the content that should be included in the IQN (See [5] Section 6.3.1). In the case of network-level prediction with PF located in the 3GPP network, such content may include the QoS KPI value, which can be either an average value, a median value, a range or Cumulative Distribution Function (CDF). [10] does not support the possibility to include the QoS KPI value in the analytics response or notification. However, the NF Consumer may specify one or more threshold(s) and NWDAF can include in the response or notification which of the threshold(s) are predicted to be crossed. A notification mechanism based on configurable thresholds has the advantage to minimize the signalling due to excessive number of notifications. However, it could be explored if V2X application layer may require also a different type of reporting which includes more information on the KPI.

Area of improvement 3

The following area of improvement has been identified:

3GPP Rel-16 Solution supports including in the notification to the consumer the prediction of the range for the QoS KPI in question according to a predefined set of thresholds, but not of the actual value of the KPI (either average or median or the CDF). While this has the advantage of minimising signalling notifications, it could be explored if it is possible to include more information on the KPI.

5GAA [5] has identified different Information Categories that could be used for QoS prediction, which includes vehicle information and client-performance measurement, RAN and CN information, and Third-Party information such as weather, coverage maps, road traffic, etc. Such Information Categories have also been detailed with examples (see [5] Section 6.2). Although 5GAA has not provided a comprehensive list of use cases where the Information Categories could be used for QoS prediction, it has provided details on how some of this information can be utilised to make application-level QoS predictions. The current 3GPP Rel-16 solution can only use OAM data as input data for prediction. While adding additional inputs to the PF does not configure automatically as an improved functionality for the consumer, it could be explored how the additional inputs could be used to further improve the quality of the prediction. Quality improvements could be related to the granularity of the prediction (as described below in Area of improvement 5), earlier detection of potential QoS changes (as described in the Area of improvement 6) as well as better accuracy and/or lower numbers of cases of prediction mismatch.

Area of improvement 4

The following area of improvement has been identified:

The 3GPP Rel-16 solution supports prediction notifications based on input data collected from OAM. Other input data, such as those from application layer (vehicle/server), RAN, CN NF and Third-Parties, could also be used for generating prediction notifications, and their usage could be investigated in order to increase the quality of the prediction.

It shall be reminded that IQN notifications may have a very real impact on the driving behaviour of a vehicle. As an example, if an IQN related to a potential QoS deterioration is received by a Tele-Operated

vehicle, the V2X application in the vehicle may have to reduce speed (sometimes even abruptly, depending on driving conditions), change lane, initiate a detour or take it to a complete stop for safety reasons. Depending on the current speed of the vehicle, the quality of driving experience may be seriously affected. If the network detects a potential QoS change that may affect a specific location, for a specific service and within a specific time window, it does not automatically mean that all of the UEs in that location accessing that specific service (and in that time window) may equally be affected. There could be factors that may impact on the actual QoS that is delivered to a PDU Session of a UE so that different UEs may actually experience very different QoS. Some of those factors include the subscription configuration, terminal specific capabilities (e.g. supported RAT types or number of antennas), and the S-NSSAI. The network shall take such information into account when generating IQNs, otherwise there is a risk that some UEs may receive a notification of potential QoS notification while their probability to actually experience the QoS change is considerably lower. As explained above, the network shall try to avoid that such type of events arise.

For these reasons, 5GAA [5] defined IQN as related to specifics of the UE and PDU Session or QoS Flow, and included in its content the relevant PDU Session Id and QoS Flow Id. Therefore, according to 5GAA, IQN is specific for a PDU Session or a QoS Flow. The IQN does not provide QoS information that relates generically to a region of the network and a specific time interval.

In 3GPP Rel-16 solution, prediction on the specific QoS KPI requested by the consumer can be requested (or subscribed to) according to Analytics Filter Information that includes the specific QoS requirements for which the prediction is requested. QoS requirements include 5QI (standardised or pre-configured), and applicable additional QoS parameters and their corresponding values (conditional, i.e. it is needed for GBR 5QIs to know the GFBR) or the QoS Characteristics attributes PDB, PER and their values. For example, consumer can request a prediction for the throughput of a Standardised 5QI=79 (V2X messages) which is a non-GBR 5QI. Consumer may also optionally include the S-NSSAI in the Analytics Filter Information.

Input data from OAM [32] used for QoS Sustainability analytics are currently RAN UE Throughput and QoS Flow Retainability. Such data are collected in the area of interest format (TAIs or Cell IDs) which is understandable by NWDAF and, as a result, cannot be differentiated according to UE characteristics, PDU Session characteristics, subscription characteristics, etc. Therefore NWDAF can only make predictions that relate to area (cell)-level. NWDAF does not collect information that allows it to generate UE-level, PDU Session-level and QoS Flow-level predictions. This means that for UEs located in the same cell and experiencing the same service (e.g. 5QI= 79 for V2X messages) NWDAF is only able to make average predictions of the relevant KPI which cannot take into account the specific context of the UE, such as subscription (e.g. silver, bronze, gold), terminal capabilities (such as RAT type, number of antennas), etc. This is simply because the collected information from OAM does not include any of those dimensions. **For this reason it is possible to conclude that the 3GPP Rel-16 solution does not support UE-level predictions.**

It has to be noted that predictions are based on collected data. Independently of the notification mechanism (e.g. individual notifications such as per UE/PDU Session/QoS Flow or group notifications) it is expected that the PF predictions will not be generated individually but rather for a group of UEs, PDU Sessions or QoS Flows that falls into specific criteria, depending on the dimensions available on the collected data. For example, if collected data may be differentiated according to RAT type and subscription type, prediction will be applicable for the group of UEs, PDU Sessions or QoS Flows that match that specific RAT type and subscription type.

As the 3GPP Rel-16 solution does not support UE-level predictions, it means that it cannot support PDU Session-level or QoS Flow-level predictions, which are even more fine-grained in terms of granularity. The QoS Flow granularity is reflected in the 3GPP Rel-16 solution as the consumer may specify the 5QI in the request or subscription for the analytics. However, the solution cannot provide notifications with

different predictions for QoS Flows with the same 5QIs of two different UEs which are under different capabilities, network and subscription conditions, regardless of whether there may be situations in which those QoS Flows may experience very different QoS, due to the different conditions. Therefore we conclude that **predictions provided by 3GPP Rel-16 solution cannot be specific for a PDU Session Id or a QoS Flow Id.**

At the same time, moving towards a finer granularity, as predictions are based on collected statistics it is important that operators make sure the measurement data set is still large enough to provide meaningful results. This may not be a trivial task when finer granularity is achieved by adding more dimensions in the analytics filters.

Area of improvement 5

The following area of improvement has been identified:

The 3GPP Rel-16 solution supports predictions on groups of UEs, PDU Sessions and QoS Flows with the granularity of 5QI, location and S-NSSAI. Predictions provided by 3GPP Rel-16 solution currently cannot be specific for a PDU Session Id or a QoS Flow Id. The granularity of 5QI, although it is supported, is intended as an average value for that 5QI computed in the location of interest and not for a specific QoS Flow identified by that 5QI in a UE's PDU Session. Finer granularity in determining the group of UEs, PDU Sessions and QoS Flows for which the prediction is applicable could be achieved if further dimensions can be added to the data that is collected, potentially also exploiting additional input data sources.

5GAA [5] defined IQN content and includes the IQN Notice Period which is defined as follows: "The time period indicating how long in advance the IQN Consumer expects to receive an IQN." How long in advance relates to the specific time when the QoS may actually change. The IQN Notice Period depends on the time that the consumer needs in order to implement the proper reactions in preparation for the potential QoS change that is predicted to happen in the near future. Such time is use-case dependent because the consumer may implement different reactions depending on which QoS change is predicted.

IQN Notice Period introduces a time requirement for the delivery of the IQN. This time requirement is use case-specific. Any prediction that cannot be delivered before the IQN Notice Period should be dropped and not delivered in an IQN because it would not be delivered on time to the consumer in order to be useful for application adaptation. This could happen because the application adaptation requires a certain time to be initiated and completed. This means that when the KPI is predicted to go below a certain threshold, the consumer needs to be notified in-advance by at least the IQN Notice Period. The consumer needs to provide the requested Notice Period for the prediction to the entity that is supposed to send the notification, so that the Prediction Function can tune when to trigger the sending of the notification. The 3GPP Rel-16 solution does not provide support for an NF consumer to specify a notice period, or a time by when the prediction is to be received by the NF consumer. However, **3GPP Rel-16 solution has specified a parameter for NWDAF in the Analytics Reporting Information, the 'Time when analytics information is needed**', as follows: If the time is reached the consumer does not need to wait for the analytics information any longer, yet the NWDAF may send an error response to the consumer, applicable to Nnwdaf_AnalyticsInfo_Request service operation. This parameter does not apply to the QoS Sustainability Analytics. It could be explored if this parameter could be applicable for the QoS Sustainability Analytics and if it could be used in place of the IQN Notice Period.

Area of improvement 6

The following area of improvement has been identified:

The 3GPP Rel-16 solution does not provide support for an NF consumer to specify a Notice Period, or a time by when the prediction notification is to be received by the NF consumer in relation to the time when the potential QoS change event is predicted.

5.6.5 Summary

With respect to the 5GAA requirements, 3GPP Rel-16 solution supports network-level prediction where predictive functionality is located in NWDAF.

According to the current 3GPP Rel-16 solution, IQN delivery from the V2X AS to the V2X application in the vehicle is out of scope of 3GPP specifications (i.e. there is no applicable 3GPP standard interface for IQN delivery to the vehicle). However, the 3GPP Rel-16 solution defines interface and procedures for IQN delivery from the 5GS to the V2X AS.

From an analysis of the current 3GPP Rel-16 solution with respect to 5GAA requirements, the following Table 5.6-3 summarises the areas of improvement that have been identified.

Table 5.6-3: Summary of identified areas of improvement for 3GPP Rel-16 Solution

Number	Area of Improvement Description
1	The 3GPP Rel-16 solution does not currently enable delivery of potential QoS change notifications to the vehicle for UE-side application adaptation. 3GPP Rel-16 Solution supports notification of potential QoS change to the V2S AS, which may share the potential QoS change notification with the UE-side of the application using user-plane which is not a 3GPP-standardised interface. Potential improvements for future 3GPP consideration include delivery of potential QoS change notifications to the vehicle (UE-side) for application adaptation according to the following options: • The usage of a modified extended NG-RAN Notification as specified by [27] in Section 5.7.2.4 and [18] Section 5.4.5.3 where NAS signalling is used to inform the UE about potential changes in the QoS parameters (i.e. 5QI, GFBR, MFBR) that the NG-RAN is currently fulfilling for the QoS Flow, where the notification to the UE is not triggered by a QoS change that has already happened but by a prediction of a potential QoS change (option 1 of Table 5.6-1). • The usage of the VAE layer to deliver predictions towards the V2X application client side (option 3 of Table 5.6-1).
2	The 3GPP Rel-16 solution does not specify predictions for the following End-to-End KPIs: • Latency for GBR or Non-GBR QoS Flows. • Packet Delivery Ratio for GBR or Non-GBR QoS Flows. • Uplink Throughput for GBR QoS Flow and Non-GBR QoS Flows (Prediction for the partial metric RAN UE Throughput is currently supported only for Non-GBR QoS Flows). • Downlink Throughput for GBR QoS Flow and Non-GBR QoS Flows (Prediction for the partial metric RAN UE Throughput is currently supported only for Non-GBR QoS Flows). • Coverage and Capability. It has to be noted that:

	• End-to-End has to be considered considering measurement points available in the 3GPP System. • For GBR QoS Flows, as long as the GBR is guaranteed, the QoS KPIs latency, packet delivery ratio, uplink throughput and downlink throughput are also guaranteed by the network. As the current 3GPP Rel-16 solution supports the prediction of the QoS Flow Retainability KPI for GBR QoS Flows, the consumer does not need a prediction for the above-mentioned KPIs for the time intervals for which the GBR is predicted to be guaranteed. A prediction of the above-mentioned KPIs could be relevant for the time intervals for which it is predicted that GBR may not be guaranteed. This would enable the consumer to know more about the type of potential QoS change that may occur. Due to the nature of V2X services, it is expected that GBR QoS Flow prediction has to be prioritised over Non-GBR QoS Flow prediction.
3	3GPP Rel-16 Solution supports including in the notification to the consumer the prediction of the range for the QoS KPI in question according to a predefined set of thresholds, but not of the actual value of the KPI (either average or median or the CDF). While this has the advantage to minimize signalling notifications, it could be explored if it is possible to include more information on the KPI.
4	3GPP Rel-16 Solution supports prediction notifications based on input data collected from OAM. Other input data such as those from application layer (vehicle/server), RAN, CN NF and third party could also be used for generating prediction notifications and their usage could be investigated in order to increase the quality of the prediction.
5	3GPP Rel-16 Solution supports predictions on group of UEs, PDU Sessions and QoS Flows with the granularity of 5QI, location and S-NSSAI. Predictions provided by 3GPP Rel-16 solution currently cannot be specific for a PDU Session Id or a QoS Flow Id. The granularity of 5QI, although it is supported, is to be intended as an average value for that 5QI computed in the location of interest and not for a specific QoS Flow identified by that 5QI in a UE's PDU Session. Finer granularity in determining the group of UEs, PDU Sessions and QoS Flows for which the prediction is applicable could be achieved if further dimensions can be added to the data that is collected, potentially also exploiting additional input data sources.
6	3GPP Rel-16 Solution does not provide support for an NF consumer to specify a notice period, or a time by when the prediction notification is to be received by the NF consumer in relation to the time when the potential QoS change event is predicted.

6 Conclusions and Recommendations

The NESQO work item proposed a number of enhancements in the 5GS aiming at introducing Predictive QoS, which is a new proactive behaviour, enabled by in-advance notifications with QoS predictions. The new functionality allows V2X applications to take care of the issues that could be caused by a potential QoS degradation in a better way, thanks to the possibility of anticipating reactions enabled by the prediction notification. This mechanism has obvious advantages compared to the reactions that could be performed after QoS degradation has already happened and been detected by the application. Consumers of In-advance QoS Notifications (IQN) can also be network functions in the 5GS; similar considerations apply when reaction to the potential QoS degradation is performed by the network, instead of an application. For Predictive QoS, NESQO defined a set of requirements, high-level procedures, message content as well as a set of proposed KPIs which could be relevant for the QoS prediction. NESQO also identified two approaches for predictions – network-level and application-level – and provided a proposal for which KPIs can be predicted by each approach.

The continued work item eNESQO has further evolved these results in two main directions:

- In the first direction, by further detailing aspects and mechanisms for making QoS predictions. In Section 5, those aspects were analysed according to network-level and application-level predictions. The same section also included an analysis of the characteristics of the two approaches (including advantages and disadvantages).
- In the second direction, by further detailing how automotive applications may take advantage of the QoS prediction. In this context, Section 4 provided examples of prediction-centric use-case descriptions with detailed logical flows of actions that may be performed by the application when an IQN is received. Section 5.5 also defined a high-level framework to be used to define application and network reactions to QoS predictions, which was further detailed in an example where the predicted QoS KPI is 'coverage'. The same section also provided a proposal for an analysis required on automotive applications in order to use QoS predictions in combination with the exiting 5GS QoS framework.

Conclusions and recommendations in these two directions are further described in the following sections.

6.1 Conclusions and Recommendations on Making QoS Predictions

For what concerns making network-level predictions, eNESQO concluded that accuracy has an important role in the QoS prediction because applications may perform different reactions depending on the level of accuracy (higher/lower/conservative) of the QoS prediction received in the notification. Moreover, the occurrence of false predictions may be mitigated by different strategies in triggering the notification: for example, while a simpler implementation could be based on the simple prediction of the KPI, a better approach could be based on taking into account not only the predicted value of KPI but also the upper/lower potential range considering accuracy. eNESQO also recommended that prediction notifications may include both the predicted value and the predicted lower bound (in the critical direction) of the KPI under analysis, as such information can be important for the consumer in order to determine the potential reaction.

eNESQO also concluded that QoS prediction can be enhanced by the prediction of specific network events. One example is the case of handover, which affects some of the predicted KPIs. For each event, the PF may generate an event description, which could be used internally by the PF in an implementation-specific manner to improve its predictions; for example, applying modifications to the

predicted QoS KPI's time function. Another example is that the PF may modify its behaviour by adapting its prediction windows according to the event description. As a third example, event description could be used to generate specific information that may be included in the IQN and provided to the consumer.

eNESQO also concluded that UE-based predictions could be a complementary solution to the network-based solution. A UE may rely on the network-based prediction for long term and on the UE-based prediction for short term. UE predictions could be used internally in the UE or also shared with other UEs, as described in Section 5.1.3.

When it comes to application-level predictions, eNESQO proposed a prediction framework using Machine Learning models based on time series and supervised learning where the PF is located in the MNO network. The framework includes forecasting of UE location based on trajectory models and of radio conditions at a precise time and location. The QoS is predicted as a mapping process that takes as input the results of these two parallel forecasting processes. The 'training' of the mapping process can be performed either with measurements taken on the vehicle devices or with collection of data from operator Network Management Systems and probes.

While application-level predictions may be achieved also with an OTT-based PF deployment, network-level predictions usually require a PF deployed in the MNO network, since the required measurements are usually only available to the MNOs. Advantages and disadvantages of MNO-based and OTT-based approaches have been discussed in Section 5.3, concluding that extreme prediction requirements, such as those of Tele-Operated driving, require PFs deployed in the MNO network. At the same time, OTT-based predictions can more easily reach global scale and have faster time-to-market because they do not require MNO interoperability and deployment of the PF in each MNO network. In order for automotive applications to benefit from network-level predictions, it is key that a fully standardised 3GPP solution is achieved. eNESQO has also described the required areas of improvements for the current 3GPP Rel-16 solution for V2X application adjustment based on QoS Sustainability Analytics in Section 5.6. It is suggested that 3GPP considers those areas of improvement in current or later 3GPP releases. Among the areas of improvement, the latter two described in Section 5.6 are of particular importance for the automotive industry, namely the achievement of a much finer granularity for the predictions, and the possibility for the prediction consumer to specify a Notice Period, or a time by when the prediction notification is to be received in relation to the time when the potential QoS change event is predicted.

6.2 Conclusions and Recommendations on the Use of QoS Prediction in Automotive Applications

eNESQO provides a "QoS prediction-centric analysis of selected use cases, namely Tele-Operated Driving, High-Definition Map Collecting, and In-Vehicle Entertainment (IVE). The analysis provided examples of event flows and of potential QoS change values (namely QoS deterioration and associated time values) when QoS prediction is made. However, the analysis was not exhaustive and it is **recommended that more detailed analysis is provided in the context of WG1, as proposed in [33].** The examples provided in this TR and the information contained in Section 5.5 provide the necessary input for further analysis.

Section 5.5 presented a structured framework to define 'reaction' to QoS prediction both on the network and application sides. The model is based on states, triggering conditions and transitions across states. Reactions are implemented during transitions. An example for use of such structured framework has also been proposed for the prediction of coverage change from *Normal Coverage* to *Coverage Enhancement*s. eNESQO **recommends that automotive applications be analysed**

according to this structured framework in order to define the states, triggering conditions and reactions that need to be performed when a transition occurs.

It is **recommended that further studies, such as the proposed work item PRESA [33], use this structured framework** and the analysis of different qualities described above in all use cases that may benefit from QoS prediction. Those include the use case in scope for NESQO and eNESQO, and potentially all of the use cases that make use of the Uu interface.

Furthermore, eNESQO **recommends the usage of QoS prediction in combination with the existing 5GS QoS framework when implementing the application adaptation**. An example of achieving such a combination can be derived from the 3GPP Framework for Live Uplink Streaming (FLUS). In this context, eNESQO concluded that it could be beneficial to **perform an analysis of different qualities for automotive applications and their relationship with experienced network conditions.** This analysis would help to define the relationships of the application with the 3GPP QoS framework (e.g. definition of target QoS boundaries associated with different application qualities) and with the QoS prediction (e.g. definition of adequate in-advance notification intervals and thresholds for prediction generation according to their impact on driving and/or vehicle behaviour and associated completion time).

Finally, eNESQO has evaluated several alternatives for QoS prediction delivery including usage of the VAE layer developed in 3GPP SA6 as a standardised solution for delivering notifications on QoS prediction to the application layer. In detail, eNESQO has considered the usage of VAE layer for supporting both server-based and client-based reactions to QoS prediction. To this aim, eNESQO has considered the exploitation of VAE layer for prediction delivery from VAE server towards the V2X application server side and has also evaluated possible enhancements for delivery towards the V2X application client side. Other alternatives for QoS prediction delivery to the V2X application (UE-side) have been considered, such as notification using NAS signalling – as in the mechanism 'Extended NG-RAN Notification to Support Alternative Service Requirements' described in [18] – which could be triggered by QoS prediction. Further studies would be required to analyse the implications and impact on NAS signalling due to this extension of notification control procedure.

Annex A: Change History

Date	Meeting	TDoc	Subject/Comment
2019-02	CC#20	A-190055	Initial draft ToC for discussion
2019-04	CC#22	A-190076	Reflecting comments received at CC#20
2019-10	CC#26	A-190203	Reflecting status after F2F#11, including updates agreed at CC#25
2019-11	CC#27	A-190249	Reflecting status after F2F#12
2019-12	CC#27	A-190251	Restructuring and removing empty sections, added introductions
2020-01	CC#28	A-200002	Corrections in section 5.1.2, to fully include the approved A-190243
2020-02	CC#29	A-200055	Including contributions approved at F2F#13

Annex B: VAE Framework

3GPP considered how to ensure the efficient use and deployment of V2X applications on 3GPP networks. To this aim, 3GPP in TS 23.286 [21] defined the V2X application enabler (VAE) layer for 4G systems together with related functional architecture, procedures and information flows. The VAE capabilities should be offered as APIs to the V2X applications. Figure illustrates the detailed V2X application layer functional model [21]. The V2X application layer functional model utilises the SEAL services as specified in 3GPP TS 23.434 [34]. In the remainder, the focus will be on the VAE layer. The utilization of VAE layer is expected to facilitate players of V2X applications (car OEMs, tiers, third parties) in interacting with 3GPP networks as it allows such players to interface with the VAE layer without the need of implementing 3GPP network functions, such as Service Capability Server (SCS) or Application Function (AF). Given its importance for adaptation and utilisation of 3GPP networks for V2X applications, the VAE framework is also considered as a deployment option in the 5GAA application layer reference architecture [35].

The *VAE server* provides the V2X application layer support functions to the *V2X application specific server* over *Vs* reference point. The VAE server interacts with the 3GPP network system over V2, MB2, xMB, Rx and T8 reference points. The EPS is considered as the 3GPP network system. The VAE server could be developed by MNOs, network vendors, or in theory by any other third party. The V2X application-specific server could be developed by car OEMs directly, car OEM's suppliers, etc. The *VAE client* provides the V2X application layer support functions to the *V2X application specific client* over *Vc* reference point. The VAE client could in theory be developed by either car OEMs directly, car OEM's suppliers, MNOs, network vendors, or any other third party according to the level of trust and security boundaries defined for the car integration. Practically speaking, the VAE client could be seen as an implementation of standardised APIs to interact with the V2X application client and the VAE server. In the VAE layer, the VAE client communicates with the VAE server over the *V1-AE* reference point, which is realised as user-plane connection via the 3GPP system between these two entities. A V1-AE message can be sent over unicast, transparent multicast via xMB, transparent multicast via MB2. In the V2X application-specific layer, the V2X application specific client communicates with V2X application specific server over *V1-APP* reference point, which is a legacy user-plane application-level connection. The VAE client of a V2X UE communicates with VAE client of another V2X UE over the *V5-AE* reference point. The V2X application-specific client of a V2X UE communicates with VAE client of another V2X-UE over the *V5-APP* reference point, which is realised e.g. by means of PC5 link. To support distributed VAE server deployments, the VAE server interacts with another VAE server over the *VAE-E* reference point.

NOTE: the functionalities of the V2X application-specific layer, V1-APP reference point and V5-APP reference point are out of scope of 3GPP.

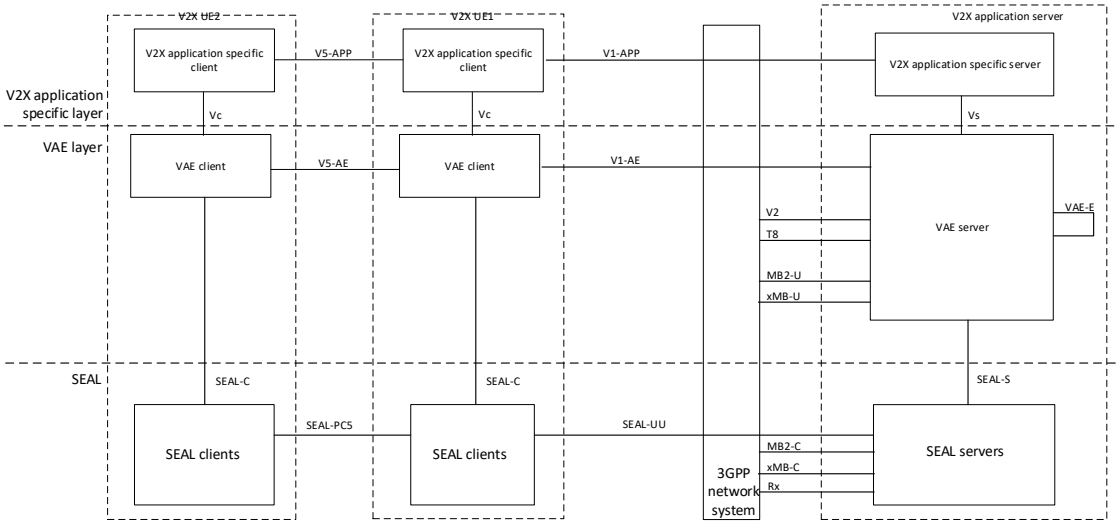


Figure B-1: V2X application layer functional model (TS 23.286 [21])

The VAE server provides the server side V2X application layer support for the following functionalities: communication with the underlying 3GPP network system (EPS) for unicast and multicast network resource management; support for registration of V2X UEs; tracking the application-level geographic location of the V2X UEs; support for V2X message distribution for the V2X applications; support for provisioning of 3GPP system configuration information (e.g. V2X USD, PC5 parameters); content provider for multicast file transfer using xMB APIs; communication of V2X service requirements to the underlying 3GPP network system (EPS); maintenance of the mapping between the V2X user ID and the V2X UE ID; support for V2X service discovery; support for V2X service continuity; and support for V2X application resource adaptation.

In addition to the above, two additional functionalities offered by VAE server are of particular interest to the QoS prediction framework. These functionalities are:

i. *Reception of monitoring reports/events from the underlying 3GPP network system* (EPS) *regarding the network situation corresponding to the RAN and core network.*
ii. *Provisioning of network monitoring reports to the V2X UEs*.

The VAE client provides the client-side V2X application-layer support for the following functionalities: registration of VAE clients for receiving V2X messages; reception of V2X messages from the VAE server and delivery to V2X application-specific client(s) according to the V2X service ID; performing the role of the MBMS client for multicast file transfer using xMB APIs; supports for switching the modes of operations for V2V communications (e.g. between direct and in-direct V2V communications); provisioning of application-level locations to the VAE server (e.g. tile, geo-fence); reception of 3GPP system configuration information (e.g. V2X USD, PC5 parameters) from the VAE server; and support for dynamic group management.

In addition to those listed above, an additional functionality offered by VAE client of particular interest to the QoS prediction framework is:

i. R*eception of network monitoring reports from the VAE server and provisioning to V2X application layer*.

As mentioned above, the VAE layer handles the provisioning of network information to V2X application layer. This procedure is detailed in [21]. The V2X UE subscribes for network-monitoring information from the VAE server and such network monitoring information may be used by the V2X UE for network connectivity adaptations. The procedure is based on a subscription request from the VAE client to the VAE server, where the request includes: V2X UE ID (i.e. identity of the V2X UE subscribing to the

network-monitoring information); subscription event (i.e. one or multiple network-monitoring events the V2X UE is interested in); and triggering criteria (i.e. identification of when the VAE server will send the monitoring reports to the VAE client). In this specification [21], the available network-monitoring information reported by the VAE server are: (i) uplink quality level; (ii) congestion level; (iii) overload level; (iv) geographical area (cell area or TA for which the monitoring applies); (v) time validity (the period for which the monitoring applies); and (vi) coverage level and bearer level events (optionally, for MBMS). The procedure for notifications for network-monitoring information is shown in Figure B-2.

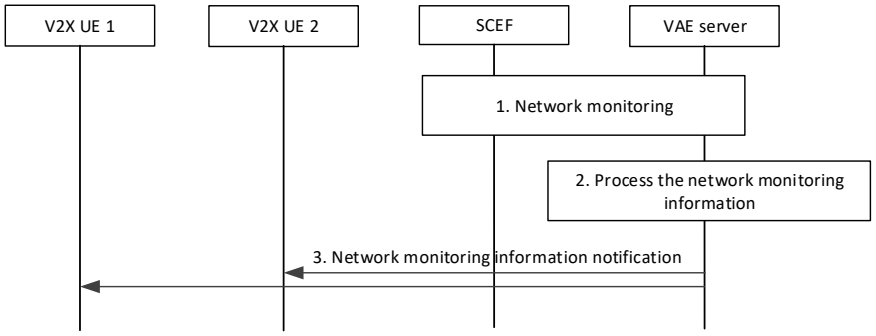


Figure B-2: Notifications for network-monitoring information (TS 23.286 [21])

The VAE server is acting as a SCS/AS which is authorised to exchange information with the Service Capability Exposure Function (SCEF) [36]. In Figure B-2, V2X UE1 and V2X UE2 have subscribed to receive network-monitoring information from the VAE server. The procedure has the following steps: the SCEF provides the VAE server with network-monitoring information [36]; and then the VAE server processes the received network-monitoring information and provides the processed network monitoring information to the subscribed V2X UEs via their respective VAE clients.

3GPP is currently working on standardization of VAE layer for 5GS (e.g., defining with which interfaces the VAE server interacts with 5G network functions). The aim is also to identify potential enhancements to the application architecture to support V2X services specified in TS 23.286 [21] and the architectural and procedural improvements for V2X services defined in TS 23.287 [18]. The outputs of this work are captured in TR 23.764 [22]. Currently, in TR 23.764 [22] the VAE server supports N33 towards 5GS (i.e. a reference point towards the NEF allowing the VAE server to gather exposed network information by 5G network functions through NEF). Among the currently open key issues, one topic considered for improving VAE-layer capabilities is related to how to support the enhancement of monitoring capabilities to be made available at the V2X application layer. This can enable the application layer to dynamically provide/adapt the service operation and related QoS requirements for a single UE or groups of UEs. The monitoring information which may be required at the application layer for adjusting the eV2X application requirement needs to be investigated further. In particular, given that the exposure of QoS information/analytics by 5GS (e.g. NWDAF) to Application Function (AF) is possible in 5GS as part of the Potential QoS change procedure (see 3GPP TS 23.287 [18]), 3GPP is currently working on enhancing the VAE layer to properly expose this notification to V2X application layer. To this aim, 3GPP is currently considering a procedure for monitoring and control of QoS for eV2X communications in [22], which is shown in Figure B-3. In this procedure, and a general pre-condition for VAE layer over 5GS, the VAE server acts as an Application Function (AF) towards the 5GS and, in detail towards NWDAF (via NEF) for this particular procedure. The VAE server subscribes to QoS monitoring services from 5GS (e.g. PCF/NWDAF). The monitoring may include the request for QoS Sustainability events, as specified in 3GPP TS 23.288 [10], which is of interest for eNESQO as considered as 3GPP baseline solution for QoS

prediction. The monitoring may include a QoS change based on Extended NG-RAN notification requests, as provided by PCF and specified in 3GPP TS 23.287 [18]. The reporting may be configured for a given area, time, periodicity etc. Based on the subscription, 5GS provides the VAE server with the desired QoS monitoring information. This report may come either from NWDAF or SMF via PCF/NEF. After processing such information, the VAE server may provide the QoS monitoring information to V2X application-specific server. The V2X application-specific server is then able to decide whether to adapt the service requirement based on received QoS monitoring information. If a QoS adaptation is required, the VAE server triggers the adaptation of QoS for the affected V2X-UE(s) by interacting with the relevant 5GS network functions. The VAE server also notifies about the QoS adaptation the V2X application specific server, which adapts the end-to-end communication towards the V2X application-specific client accordingly.

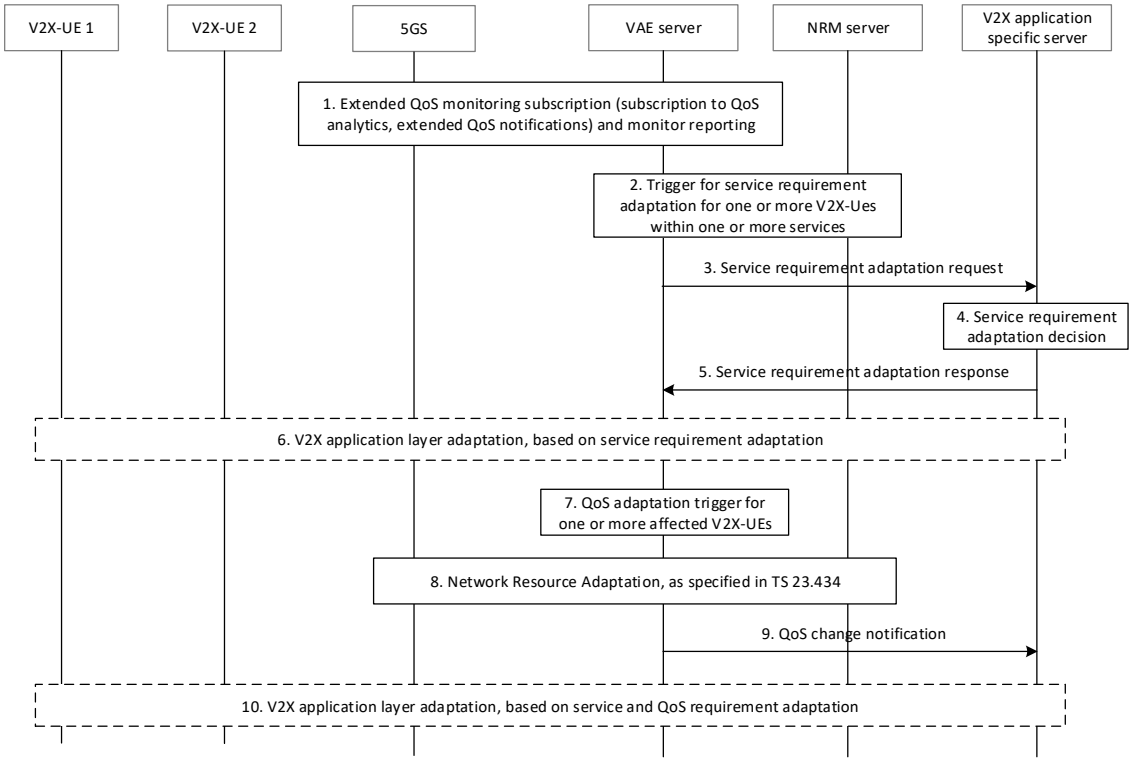


Figure B-3: Monitoring and control of QoS for eV2X communications (TR 23.764 [22])