ETSI TR 103 702 V1.2.1 (2024-08)

Speech and multimedia Transmission Quality (STQ); QoS parameters and test scenarios for assessing network capabilities in 5G performance measurements

Reference
RTR/STQ-00240m

Keywords
5G, data, LTE, LTE-Advanced, measurement, performance, QoE, QoS, service, test

ETSI

650 Route des Lucioles
F-06921 Sophia Antipolis Cedex - FRANCE

Tel.: +33 4 92 94 42 00 Fax: +33 4 93 65 47 16

Siret N° 348 623 562 00017 - APE 7112B
Association à but non lucratif enregistrée à la
Sous-Préfecture de Grasse (06) N° w061004871

Contents

Intellectual Property Rights

Essential patents

IPRs essential or potentially essential to normative deliverables may have been declared to ETSI. The declarations pertaining to these essential IPRs, if any, are publicly available for **ETSI members and non-members**, and can be found in ETSI SR 000 314: *"Intellectual Property Rights (IPRs); Essential, or potentially Essential, IPRs notified to ETSI in respect of ETSI standards"*, which is available from the ETSI Secretariat. Latest updates are available on the ETSI Web server (https://ipr.etsi.org/).

Pursuant to the ETSI Directives including the ETSI IPR Policy, no investigation regarding the essentiality of IPRs, including IPR searches, has been carried out by ETSI. No guarantee can be given as to the existence of other IPRs not referenced in ETSI SR 000 314 (or the updates on the ETSI Web server) which are, or may be, or may become, essential to the present document.

Trademarks

The present document may include trademarks and/or tradenames which are asserted and/or registered by their owners. ETSI claims no ownership of these except for any which are indicated as being the property of ETSI, and conveys no right to use or reproduce any trademark and/or tradename. Mention of those trademarks in the present document does not constitute an endorsement by ETSI of products, services or organizations associated with those trademarks.

DECT™, **PLUGTESTS™**, **UMTS™** and the ETSI logo are trademarks of ETSI registered for the benefit of its Members. **3GPP™** and **LTE™** are trademarks of ETSI registered for the benefit of its Members and of the 3GPP Organizational Partners. **oneM2M™** logo is a trademark of ETSI registered for the benefit of its Members and of the oneM2M Partners. **GSM**® and the GSM logo are trademarks registered and owned by the GSM Association.

Foreword

This Technical Report (TR) has been produced by ETSI Technical Committee Speech and multimedia Transmission Quality (STQ).

Modal verbs terminology

In the present document "**should**", "**should not**", "**may**", "**need not**", "**will**", "**will not**", "**can**" and "**cannot**" are to be interpreted as described in clause 3.2 of the ETSI Drafting Rules (Verbal forms for the expression of provisions).

"**must**" and "**must not**" are **NOT** allowed in ETSI deliverables except when used in direct citation.

Introduction

As the industry deploys 5G networks and launches 5G plans and services, it is acknowledged that it will facilitate improved service experience and the enablement of new business and services. This will include a variety of services, applications and use cases ranging from those requiring high data rates through enhanced Mobile Broadband (eMBB) to those requiring ultra-Reliable Low Latency (uRLLC) as well as those supporting massive Machine Type Communication (mMTC). Many existing applications and services such as voice, data and video will continue to be widely used with the expectation that they will benefit through superior quality, reduced access times and greater reliability. In addition, new use cases, applications and service scenarios, which are facilitated by 5G, will have specific performance requirements that require measurement and evaluation. As operators develop 5G service strategies, establish network requirements and develop networks to meet those requirements, it is important to be able to quantify and qualify the capabilities of the network. To achieve this, it is necessary to examine what QoS parameters should be measured to quantify a networks capability and how the resultant service or application QoS will be assessed. To that end, the purpose of the present document is to identify those QoS parameters and the test scenarios that can evaluate 5G performance and measure the network capability and readiness.

At the current stage of network development, with the focus on data rates, the present document will focus on eMBB and the use cases and services it enables such as ultra-high definition video or virtual reality as primary examples. It should be noted that while the performance requirements will evolve during the lifecycle of the 5G network to meet new use cases and customer expectations, the QoS parameters and test scenarios will continue to provide a means to measure the network capabilities to meet these requirements.

There are many aspects of advanced technology and 5G features to consider when examining the impact on quality of service including MEC, where interactivity requirements influence deployment strategies, to network slicing, where context aware intelligence directs traffic according to application requirements, to radio features such as beam forming and massive MIMO, which provide intelligent management of the air interface and many others. Given the complexity of 5G features, the aim of the present document is to focus on the end to end network capability with reference to 5G feature considerations where required. In this regard, it is necessary to examine performance measurement scenarios [i.2], to determine the network technical parameters and consider the impact on end user service experience [i.6]. In addition, to support QoS parameter measurement, the test scenarios that measure the network capability will be described.

1 Scope

Given the current stage of 5G deployments and the focus on eMBB, the present document, will primarily concentrate on QoS parameters in relation to eMBB performance scenarios and the most prevalent eMBB related service scenarios. Therefore, the focus will be on QoS parameters that reflect the network capabilities in the case of visual applications such as UHD video and Virtual and Augmented Reality. However, whilst these technologies, are primarily categorized as requiring high data rates, it is necessary to be mindful and examine the relationship and requirements in high data rates and low latency scenarios.

In the scope of this analysis, the term QoS relies on service-related characteristics without knowing any details about the underlying network sections [i.6], the network architecture and the network or application deployment strategies. The scope will concentrate on measuring the network capability, assessed primarily through network QoS parameters, such as data rates, capacity, coverage, latency and continuity measurements. The readiness of the network to support the QoS needs of existing services and applications such as voice, data and video and those newly enabled technologies and use cases mentioned earlier that benefit from the higher data rates associated with eMBB will also be considered.

The approach therefore will be to assess network capabilities by first identifying the performance scenario requirements, then discovering the QoS parameters that will measure those requirements and finally defining the test scenarios to measure those QoS parameters as follows:

1) Identify scenarios in terms of performance, service and user types to determine the performance measurement requirements and the key performance factors that will satisfy those requirements:

   - Performance scenarios [i.2] which are dependent on traffic types, traffic densities and service areas.

   - Service scenarios that consider the use cases, technology and applications that place data service requirements on the network for effective operation.

   - User type scenarios that examine various types of users and how they place different requirements on similar services.

2) QoS parameter discovery to identify and define the parameters that represent the key performance factors and scenario requirements. The QoS parameters will define how to effectively measure the network technical performance as well as examining how a use case or application might be affected by those network conditions. The QoS parameters in as much as possible will refer to existing definitions and best practises.

3) Provide test scenario analysis to detail the types of tests to be executed to verify the network capability. Define how to represent the measurement scenarios and where to collect data to calculate the QoS parameters. The test scenarios will reproduce typical service activities to derive quality measures and will identify the measurement points and the expected data sources.

The aim therefore is to identify the QoS parameters of interest, based on the identified scenarios, referencing existing specifications and technical reports where available. There are already significant relevant references available from a number of bodies to identify QoS aspects for a number 5G scenarios, which are at various stages of maturity. This includes analysis of primary use case scenarios, identification of performance measurement scenarios and definitions of quality measurement indicators. The expectation is that, the present document, through its analysis will put in place a means to assess 5G network capabilities and readiness of the network to support those aforementioned prevalent eMBB applications and use cases.

2 References

2.1 Normative references

Normative references are not applicable in the present document.

2.2 Informative references

References are either specific (identified by date of publication and/or edition number or version number) or non-specific. For specific references, only the cited version applies. For non-specific references, the latest version of the referenced document (including any amendments) applies.

NOTE: While any hyperlinks included in this clause were valid at the time of publication, ETSI cannot guarantee their long term validity.

The following referenced documents are not necessary for the application of the present document but they assist the user with regard to a particular subject area.

[i.1] ETSI TS 102 250-2: "Speech and multimedia Transmission Quality (STQ); QoS aspects for popular services in mobile networks; Part 2: Definition of Quality of Service parameters and their computation".

[i.2] ETSI TS 122 261: "5G; Service requirements for the 5G system (3GPP TS 22.261)".

[i.3] ETSI TR 101 578: "Speech and multimedia Transmission Quality (STQ); QoS aspects of TCP-based video services like YouTube™".

[i.4] ETSI TR 126 918: "Universal Mobile Telecommunications System (UMTS); LTE; Virtual Reality (VR) media services over 3GPP (3GPP TR 26.918)".

[i.5] ETSI TR 126 929: "5G; QoE parameters and metrics relevant to the Virtual Reality (VR) user experience (3GPP TR 26.929)".

[i.6] ETSI TS 102 250-1: "Speech and multimedia Transmission Quality (STQ); QoS aspects for popular services in mobile networks; Part 1: Assessment of Quality of Service".

[i.7] Recommendation ITU-T G.QOE-VR: "Influencing Factors on Quality of Experience (QoE) for Virtual Reality Services".

[i.8] Recommendation ITU-R M.2083-0: "IMT Vision - Framework and overall objectives of the future development of IMT for 2020 and beyond".

[i.9] ETSI TS 103 222-2: "Speech and multimedia Transmission Quality (STQ); Reference benchmarking, background traffic profiles and KPIs; Part 2: Reference benchmarking and KPIs for High speed internet".

[i.10] ETSI TS 102 250-3: "Speech and multimedia Transmission Quality (STQ); QoS aspects for popular services in mobile networks; Part 3: Typical Procedures for Quality of Service measurement equipment".

[i.11] ETSI TS 102 250-5: "Speech and multimedia Transmission Quality (STQ); QoS aspects for popular services in mobile networks; Part 5: Definition of typical measurement profiles".

[i.12] ETSI TR 103 468: "Speech and multimedia Transmission Quality (STQ); Quality of Service aspects for 5G; Discussion of QoS aspects of services related to the 5G ecosystem".

[i.13] "5G Service Experience-Based Network Planning Criteria", Ovum in partnership with Huawei, Copyright Ovum 2019©.

[i.14] 3GPP TS 26.186: "Enhancement of 3GPP support for V2X scenarios; Stage 1; Release 16".

[i.15] 3GPP TR 29.893: "Study on IETF QUIC Transport for 5GC Service Based Interfaces (Release 16)".

[i.16] Recommendation ITU-T Y.1540: "Internet protocol data communication service - IP packet transfer and availability performance parameters".

[i.17] IETF RFC 5357: "A Two-Way Active Measurement Protocol (TWAMP)".

[i.18] Recommendation ITU-T P.1204: "Video quality assessment of streaming services over reliable transport for resolutions up to 4K".

[i.19] ETSI TR 103 559: "Speech and multimedia Transmission Quality (STQ); Best practices for robust network QoS benchmark testing and scoring".

[i.20] IETF RFC 5481: "Packet Delay Variation Applicability Statement".

[i.21] IETF RFC 6038: "Two-Way Active Measurement Protocol (TWAMP) reflects Octets and symmetrical size features".

[i.22] ETSI TS 123 501: "5G; System architecture for the 5G System (5GS) (3GPP TS 23.501 Release 16)".

[i.23] Recommendation ITU-T G.1051: "Latency measurement and interactivity scoring under real application data traffic patterns".

3 Definition of terms, symbols and abbreviations

3.1 Terms

Void.

3.2 Symbols

Void.

3.3 Abbreviations

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

AI	Artificial Intelligence
AR	Augmented Reality
CSI	Channel State Information
DL	DownLink
DNS	Domain Name System
eMBB	enhanced Mobile BroadBand
EN-DC	E-UTRAN New Radio Dual Connectivity
FOV	Field Of View
HD	High Definition
HTML	Hyper-Text Meta Language
HTTP	HyperText Transfer Protocol
HTTPS	HyperText Transfer Protocol Secure
ICMP	Internet Control Message Protocol
IETF	Internet Engineering Task Force
IMAX	Image MAXimum
IMT	International Mobile Telecommunications
IoT	Internet of Things
IP	Internet Protocol
IPDV	Inter-Packet Delay Variation
ITU-T	International Telecommunication Union - Telecommunication Standardization Sector
KPI	Key Performance Indicator
LQO	Listening Quality Objective
MEC	Multi-Access Edge Computing
MIMO	Multiple Input Multiple Output
ML	Machine Learning
mMTC	massive Machine Type Communication
MOS	Mean Opinion Score
MTP	Motion to Photon
NR	New Radio
OSS	Operations Support System

OTT	Over The Top
PDV	Packet Delay Variation
QoE	Quality of Experience
QoS	Quality of Service
QUIC	Quick UDP Internet Connections
RSCP	Received Signal Code Power
RSRP	Reference Signal Receive Power
RTP	Real-Time Protocol
RTT	Round Trip Time
SDK	Software Development Kit
SINR	Signal to Interference plus Noise Ratio
TCP	Transmission Control Protocol
TLS	Transport Layer Security
TV	Tele-Vision
TWAMP	Two Way Active Measurement Protocol
UDP	User Datagram Protocol
UE	User Equipment
UHD	Ultra High Definition
UL	UpLink
URL	Uniform Resource Locator
URLLC	Ultra Reliable Low Latency Connection
V2X	Vehicle to X (anything)
VoNR	Voice over New Radio
VR	Virtual Reality

4 5G Performance measurement criteria

4.1 Overview

5G is an evolution of existing mobile technologies, which initially leverages existing LTE networks through EN-DC and on towards NR deployments. The types of scenarios which require higher data rates through eMBB, reliable low latencies through uRLLC and the low energy, high coverage associated with mMTC will each have their own performance characteristics.

IMT 2020, envisions a broad variety of capabilities, tightly coupled with intended usage scenarios and applications [i.8]. The intention being that for different usage scenarios, these capabilities will have varying degrees of relevance and significance. Therefore, this clause identifies the performance measurement criteria for the identified scenarios.

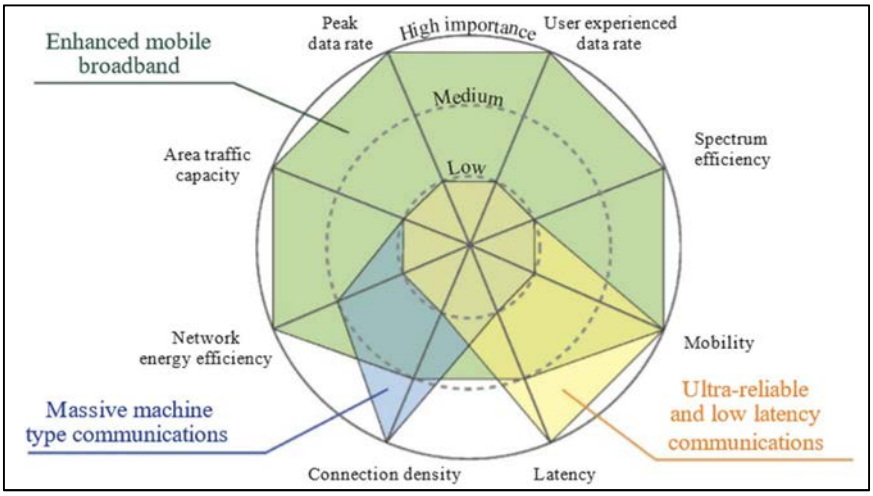

Figure 1: The importance of key capabilities in different usage scenarios

The 5G system is expected to provide optimized support for a variety of different services, different traffic loads and different end user communities [i.2]. In this regard, 5G performance measurement requirements are examined based on the expected scenarios, which can be categorized as:

- Traffic Scenarios.
- Service Usage Scenarios.
- User Type Scenarios.

The parameters to measure the performance to meet the needs of the discovered scenarios are identified, along with the expected and best practise performance levels.

4.2 Traffic Scenarios

4.2.1 Overview

Performance is highly dependent on traffic scenarios, which are identified for specific service areas e.g. urban and rural, where an urban scenario, needs to provide high data rates and high capacity whereas for a rural scenario, the main priority is to provide coverage with a minimum useful data rate. This will support services such as high definition video, cloud gaming and virtual reality in an urban setting whereas browsing and video may be more important in a rural environment.

4.2.2 Key capability parameters

The following parameters are considered to be key capabilities, within the scope of the present document, with regard to performance of traffic scenarios:

- **User experienced data rate:** achievable data rate that is available across the coverage area to a mobile user/device (in Mbit/s or Gbit/s) at the application layer.
- **Peak Data Rate:** maximum achievable data rate under ideal conditions (in Mbit/s or Gbit/s).
- **Latency:** the time from when the source sends a packet to when the destination receives it (in ms).
- **Mobility:** maximum speed at which a defined QoS can be achieved (in km/h).
- **Area traffic capacity:** total traffic throughput served per geographic area (in Mbit/s/m^2).
- **Coverage:** in this instance, defined as network coverage, which is the total land area covered by 5G signal divided by total land area.

4.2.3 High data rates and traffic densities

Scenarios which need high data rates and traffic densities as illustrated in Figure 2, demand, high UL and DL traffic capacity (50 - 100 Gbps/km^2) and high user experienced UL and DL data rates (25 - 50 Mbps) as identified in clause A.1 of the present document. Coverage requirements range from full network in urban and rural regions to specific areas in Indoor and Dense Urban regions and along traffic routes such as roads and railways. Mobility requirements range from pedestrians to high speed vehicles, trains and aircraft.

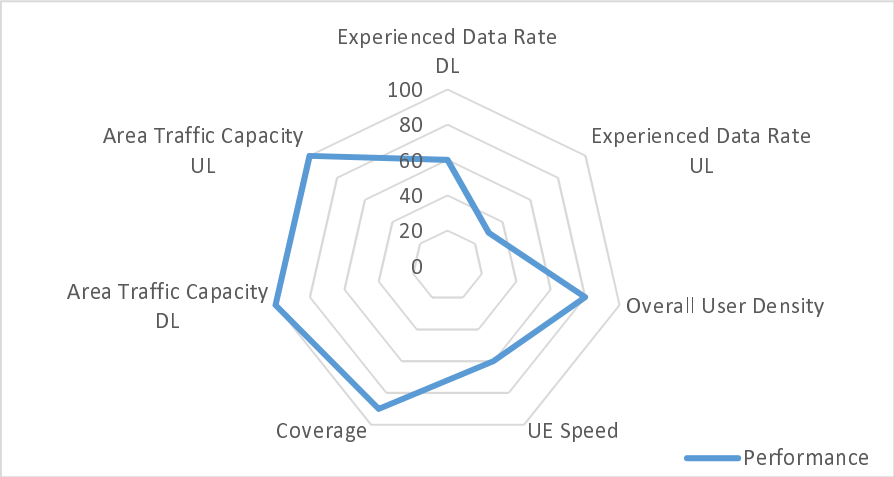

Figure 2: High data rate and traffic density performance capabilities

Several scenarios need the support of very high data rates or traffic densities of the 5G system [i.2]:

- Urban macro - The general wide-area scenario in urban area.
- Rural macro - The general wide-area scenario in rural area.
- Indoor hotspot - The scenario for offices and homes, and residential deployments.
- Broadband access in a crowd - The scenario for very dense crowds, for example, at stadiums or concerts.
- Dense urban - The scenario for pedestrian users, and users in urban vehicles, for example, in offices, city centres, shopping centres, and residential areas.
- Broadcast-like services - The scenario for stationary users, pedestrian users, and users in vehicles, for example, in offices, city centres, shopping centres, residential areas, rural areas and in high speed trains.
- High-speed train - The scenario for users in trains.
- High-speed vehicle - The scenario for users in road vehicles.
- Airplanes connectivity - The scenario for users in airplanes.

4.2.4 High data rate and low latency

Scenarios which need high data rates and low latencies as illustrated in Figure 3, demand, maximum allowed end to end latencies (5 - 10 ms) and relatively high data rates (100 Mbps - 1 Gbps) as identified in clause A.1 of the present document. Coverage requirements range from countrywide to small geographical areas. Due to the nature of use cases reliant on this type of scenario the mobility requirements are primarily stationary or pedestrian. This scenario has reliability requirements in the uplink (99,90 %) and downlink direction (99,9 %) and support for a relatively small number of UEs (< 10). The end to end latency depends not only on the connectivity delay which includes the radio interface and network transmission but also delays which may be outside the 5G system.

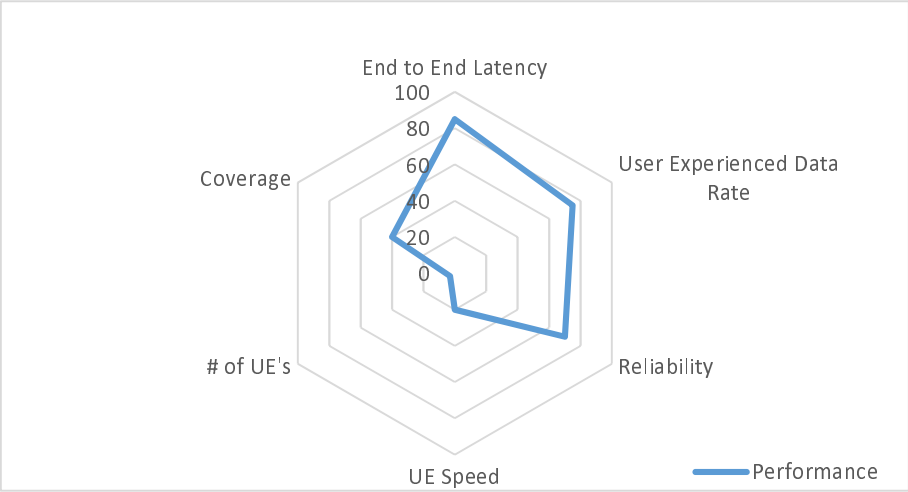

Figure 3: High data rate low latency performance capabilities

Several interactive services need the support of very high data rates, and low latency [i.2]:

- Cloud/Edge/Split Rendering - characterized by transmitting and exchanging the rendering data between the rendering server and device.
- Gaming or Training Data Exchanging - characterized by exchanging the gaming or training service data between two AR/VR devices.
- Consume VR content via tethered VR headset - tethered VR headset receiving VR content via a connected UE.

The type of use cases and technologies supported by these performance criteria include Augmented Reality, Virtual Reality, Gaming and Training. Audio-visual interaction defined in ETSI TS 122 261 [i.2], focuses on the requirements for audio-visual feedback where the VR environment interaction requires the 5G system to support low motion-to-photon capabilities from the physical movement of a user's head to the updated picture in the VR headset.

4.3 Service usage scenarios

4.3.1 Overview

Performance criteria vary by service or application where for example, streaming video or VR 360 have different network needs to more interactive applications such as fully immersive virtual reality or online gaming. This necessitates understanding how to measure the QoS in relation to a network performance capabilities and an operator's service strategy.

The service usage scenarios examine performance criteria for existing services and newly enabled services, technologies and applications. Use cases are being driven by the increased demand for data, particularly through streaming services which support video, music, user generated content and highly interactive experience such as VR and AR. These scenarios place rigorous requirements on the network, increasingly with real time needs, for higher date rates and lower latency in particular. The aim is to determine the performance criteria for each of the usage scenario categories .

To fully understand the performance criteria of the many different scenarios, necessitates identification of the technologies and the services that will be supported. The approach is to identify categories of these scenarios and determine common requirements. An example illustrates this approach, examining remote education (identification of data rate, latency and mobility performance criteria):

- AR Assisted teaching - supported by AR & UHD Video services with performance criteria for Up/Downlink Rates, Latency, Location Accuracy, Mobility.
- VR Assisted teaching - supported by VR services with performance criteria for Up/Downlink Rates, Latency.

The example indicates how a scenario may place varying performance requirements for data rates, latency, mobility and others on the network depending on the technology, application and service necessary to deliver the use case. The approach requires analysis of the technology and categorization of the different types of way that the applications can be consumed by users and identify the performance criteria for these categories. To explore this further, in the above example, depending on the offering, there are different categories of VR applications which place different requirements on the network depending on whether they are highly interactive or 360 degree video VR applications. It is important to distinguish the performance criteria for these scenarios in categories so they can be effectively evaluated. The clause in general is examined from two viewpoints:

- Performance criteria for existing services influenced by greater experience expectations.
- Performance criteria for newly enabled usage scenarios, technologies and services.

While data reliant scenarios place performance requirements on the network, these are reflected from the user's perspective in terms of the impact on the service or application. For these usage scenarios it is necessary for the QoS to measure and assess the ability to access and retain the service connection, the time taken to access, the quality of the delivery and the interactive capability based on the assessment of the overall transmission chain from a user's perspective [i.6].

4.3.2 Key service usage scenario performance parameters

The following parameters are considered to be key capabilities, with regard to performance of usage scenarios:

- **IP-Service Access:** characterizes the time needed to initialize and start the target application. It covers the DNS resolution as well the time until the actual transfer of payload data begins. Based on a time-out for the IP-Service Access time, a failure ratio or success ratio respectively can be derived.
- **Data Rate:** characterizes the transmission speed, or the number of bits per second transferred and is a key performance measurement to differentiate 5G, especially in scenarios supporting enhanced mobile broadband.
- **Interactive Latency or Delay:** represents two way latency or RTT of the QoS delay experienced at the user terminal from requesting a service or performing an interactive action to receiving the appropriate response to establish the service or provide the interactive response. Interactive latency or delay includes delays in the terminal, network, and any servers.
- **Delay Variance:** delay variation relates to the variability in arrival times of individual packets. Services that are highly intolerant of delay variation will usually take steps to remove (or at least significantly reduce) the delay variation by means of buffering, but in the case of real time scenarios delay variance has a significant effect.
- **Information Loss:** has a direct effect on the quality of the information presented to the user or for the further usage by e.g. a production chain or a machine, whether it is voice, image, video or data. In this context, information loss is not limited to the effects of packet loss during transmission, but also includes packets that arrived too late and the effects of any degradation such as corrupt packets on arrival, in other words packet corruption. Information loss is not only caused by transmission, compression artefacts and non-optimal de-compression also lead to information loss. Information loss or quality of the presentation is preferably measured by integrative measures weighting the individual information losses and distortions correctly to each other, considering cross-masking effects and provide an integrative metric.

It should be recognized that these performance factors, whilst measuring network capability have a significant impact on the scenario QoS, where for example low data rates and delays have considerable effect in terms of accessibility, retainability and quality.

4.3.3 Existing services & applications

The consideration of performance criteria for existing services indicate that the deployment of 5G will impact on existing services through:

- significantly higher download data rates;
- increased upload data rates;
- higher reliability through better success rates in terms of accessibility to services;

- shorter access times, meaning quicker times to access a service;
- lower stalling, less freezing leading to better quality;
- more spatial and temporal details by higher resolution and frame rates and less compression along with more reliable (no packet loss) transmission, meaning better quality;
- lower delay and lower delay variation will lead to short buffer-times and therefore to lower lag to real-time (e.g. live video).

It is expected that usage scenarios will be similar to existing use, but there will be a greater propensity for upload and real time in particular with regard to social media information.

The performance criteria for Voice service will not change significantly initially but will need re-evaluation in terms of VoNR voice service which is expected to support lower call set up times and higher voice quality.

There will be a greater impact on data services including download and upload of significantly increased data volumes and in scenarios such as web browsing, video, messaging and social media interaction. The network performance criteria will necessitate QoS parameters to evaluate increased data rates and reduced latency resulting in higher accessibility and retainability success rates and greater quality in terms of response times and presented resolution. Web browsing, social media, messaging and video streaming involve uploading and downloading content from a server, meaning that bandwidth or data rate is the major factor. Latency is important in relation to upload and sharing of real time content, with higher latencies impacting the quality of real time social media video streaming including entertainment, conversational video and video conferencing.

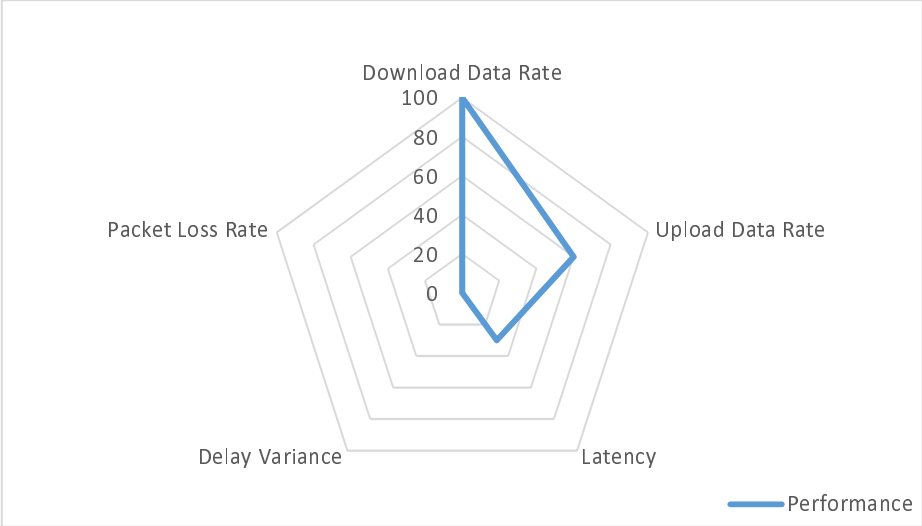

Figure 4: Existing data services performance capabilities

The following performance measurements and associated criteria will be priority for existing data services:

- Data rate (UL) (Mbps).
- Data rate (DL) (Mbps).
- Latency/Delay (ms).

4.3.4 5G enabled services, applications & technologies

4.3.4.1 Overview

Many new services and applications will be enabled through 5G, but it is clear that technologies that require higher data rates are at the forefront of the development and enablement of new usage scenarios. This includes enhanced video, virtual reality and the continuing development opportunities in online gaming, where graphics are becoming more lifelike and interactions more real time. The aim therefore is to identify the usage scenarios and the performance criteria for scenarios and applications that deliver visual information.

4.3.4.2 Enhanced video streaming

It is understood that the performance criteria for video will be based on higher resolutions, such as UHD-1 ('4K') and UHD-2 ('8K') under the condition of delivering more spatial details, shorter access times to video services and finally better quality in terms of video MOS, freezing and stalling. UL and DL data rates are necessary to ensure the defined resolution quality is achieved. Download has the greatest impact on video streaming but with social media sharing, real time uploads and latency will also be important. The criteria are identified in clause A.2 of the present document. For video streaming short latencies enable less lag to real-time video transmissions. There is a clear tendency that video streaming is moved to reliable transmission by TCP or increasing other securing mechanisms as e.g. QUIC. Therefore, sporadic packet loss becomes less important for video streaming:

- Data rate (DL) (Mbps).
- Data rate (UL) (Mbps).
- Latency/Delay (ms).

4.3.4.3 Enhanced Video Conferencing

The same as for video streaming, also video conferencing or video calls will enhance by higher resolution and frame-rates (delivering more details, meaning there is less information loss to the original video signal). Compared to video streaming, video conferencing and stronger demands on Data rate (UL) because of sending a video signal in uplink and especially for latency to ensure a real-time experience while conferencing. Video conferencing require very short delays, there is almost no possibility to buffer video frames to compensate delay variations and Delay Variance becomes a QoS parameter in this case. Interactive visual applications as video conferencing based usually on non-reliable transmission; therefore, packet loss is an important QoS parameter too:

- Data rate (DL) (Mbps).
- Data rate (UL) (Mbps).
- Latency/Delay (ms).
- Delay Variance (ms).
- Packet Loss (%).

4.3.4.4 Messaging & Visual communication

Messaging and in particular OTT messaging will become more real time, interactive and be will be primarily based on video communication. The growth of video communication as a conversational tool has always been important but it has become critical in the face of worldwide pandemics. A lot of this messaging and communication occurs at home where wireline connectivity may be prominent but in terms of faster speeds, increased quality and reliability 5G offers a significant alternative in terms of fixed wireless access or whilst on the move through mobile data. Similarly to video streaming, short latencies are important for real-time chat and video transmissions. This along with higher throughput capabilities enable the transfer of larger data volumes sharing pictures, audio and video clips:

- Data rate (DL) (Mbps).
- Data rate (UL) (Mbps).
- Latency/Delay (ms).

4.3.4.5 Virtual Reality

The relevance of Virtual Reality and its possible points of interoperability including potential for standardization are examined in ETSI TR 126 918 [i.4]. In particular relevant to the present document is the network impact of latency and delays discussed in clause 9.4.2 of ETSI TR 126 918 [i.4]. Additionally, there is significant analysis of Virtual Reality QoS, which includes identification of use cases [i.4], quality factors [i.5], [i.7] and other considerations. The focus of this analysis is the end to end VR service including the terminal, the network and the content. The analysis in the present document is focused on the network capabilities and identification of the network performance criteria that impact on virtual reality usage scenarios. Analysis of these VR scenarios indicate that they fall into some basic categories.

Cloud Virtual Reality video type applications are a semi immersive type of virtual reality, typically delivered via the cloud as VR panoramic video, VR live broadcast or VR 360 degree videos. It is the immersive experience of being able to view in every direction at the same time and the view port can be delivered as a full view 360 degree download or rendered based on the user's field of view. VR video is delivered in a manner very similar to video streaming, so as the user watches the content, they can view the content typically in 360 degrees. This can be delivered across the network in 2 ways:

1) the 360° view is downloaded to the device and streamed as a video the content is played. Typically this is delivered in high definition requiring substantial data rates;

2) the Field Of View (FOV) relevant to the current view port is rendered and delivered to the device according to the movement of the users head. Any delay in rendering viewports is a critical QoS factor.

The cloud VR video chart below, illustrates the dominance of data requirements for download of the view either as a buffered video download or the newly rendered scenes according to the users movement. There is no upload component as there is no interaction with the environment.

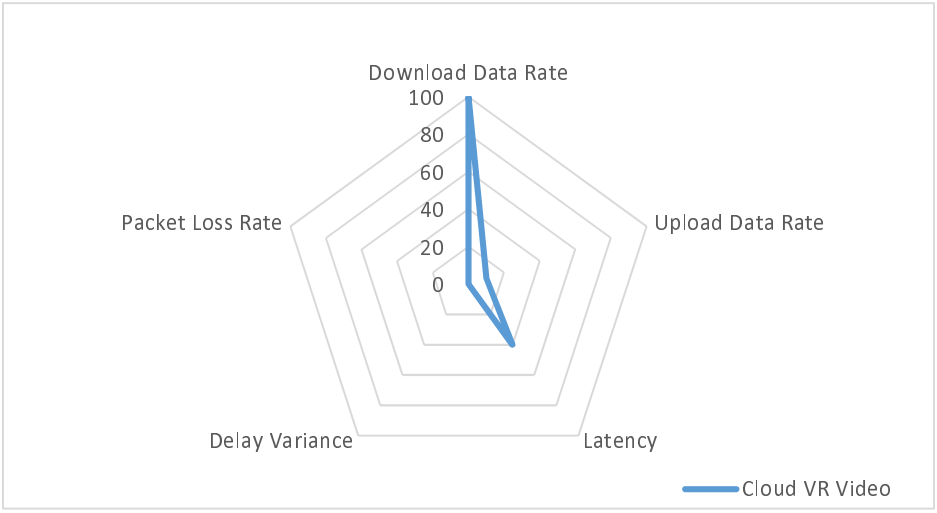

Figure 5: Cloud VR Video performance capabilities

The performance criteria for cloud VR video categorized services can be identified by the following QoS parameters. Network requirements are identified in clause A.2 of the present document:

- Data Rate (DL) (Mbps).
- Latency/Delay (ms).
- Delay Variance (ms).

Cloud Virtual Reality interactive type applications create a fully immersive virtual reality experience where the user has control of their environment beyond direction. Instead of just a viewport, a user can move virtually and interact with the environment. Objects can be picked up, doors can be opened and other players can be chased in an interactive manner and so on. The key differences to be measured from a QoS point of view is that the view is rendered according to the movement and interaction within the environment. Typical scenarios include VR Games, VR Education, VR Home Fitness and VR Social Networking.

The cloud VR interactive chart, illustrates the dominance of data and latency requirements for user interaction in a VR environment that responds to users motion, interacts in real time, without the effect caused by delays, poor visual quality or motion sickness. The data rate needs to be guaranteed to meet the content resolution requirements and the delay and consistency of the connectivity sufficient to ensure that the user's movements are interpreted with a seamless response.

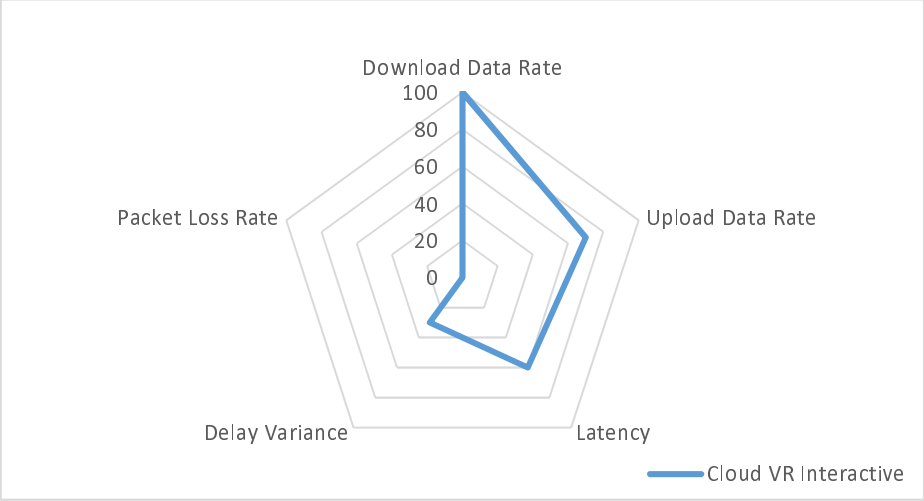

Figure 6: Cloud VR Interactive performance capabilities

The performance criteria for cloud VR interactive categorized services can be identified by the following QoS parameters. Network requirements are identified in clause A.2 of the present document:

- Data Rate (DL) (Mbps).
- Data Rate (UL) (Mbps).
- Latency/Delay (ms).
- Delay Variance (ms).
- Packet Loss (%)

4.3.4.6 Cloud gaming

Cloud gaming is similar to VR in terms of QoS in relation to audio-visual quality, interaction quality, and the consistency and integrity of the audio-visual experience. For cloud gaming, the game picture is processed and rendered on the cloud, compressed and the rendered view transmitted to clients through the network. Depending on the game, at a high level this can be categorized, as those which deliver high graphical content i.e. requiring high data rate from a network capability point of view and those that are highly interactive i.e. requiring low latency. There are games that are highly graphical and also highly interactive.

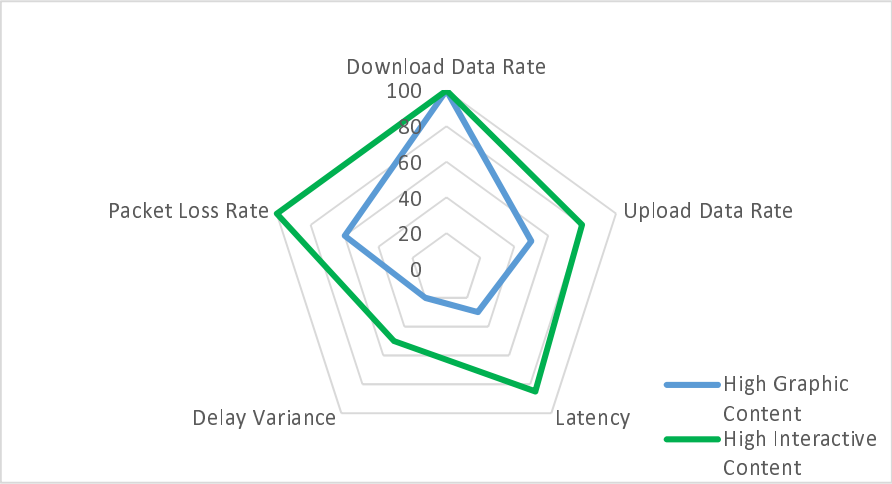

Figure 7: Cloud gaming performance capabilities

Typically, responsiveness is the most critical user-perceived QoS measurement for cloud gaming systems. This responsiveness can be described as the time duration between a user submitting a command and the time the corresponding game frame is displayed back to the user and directly affects the user's performance and gaming experience.

Cloud Gaming performance criteria are similar to interactive virtual reality requirements, although there is a greater emphasis on latency and a larger emphasis on data rates for the highly graphical type of games. The performance criteria for cloud gaming can be identified by the following QoS parameters. Network requirements are identified in clause A.2 of the present document:

- Data Rate (DL) (Mbps).
- Data Rate (UL) (Mbps).
- Latency/Delay (ms).
- Packet Loss (%).
- Delay Variance (ms).

4.3.4.7 Augmented Reality

Augmented Reality (AR) as an interactive experience of a real-world environment where the objects that reside in the real world are enhanced by computer-generated perceptual information requires high data rates, low latency and location accuracy. The upload data rates for AR sensor is significant to provide context for the real-world scenario and can be as high as or higher than the downlink requirements. The latency requirement for AR scenarios are higher (meaning lower latency is required) than VR as visual changes are not only triggered by the motion of the user but also by any change (e.g. lighting or natural object movement) in the surrounding world. The location or position of the device is important as it is necessary to enrich the item in the frame with additional information in scenarios such as tourism, manufacturing or gaming.

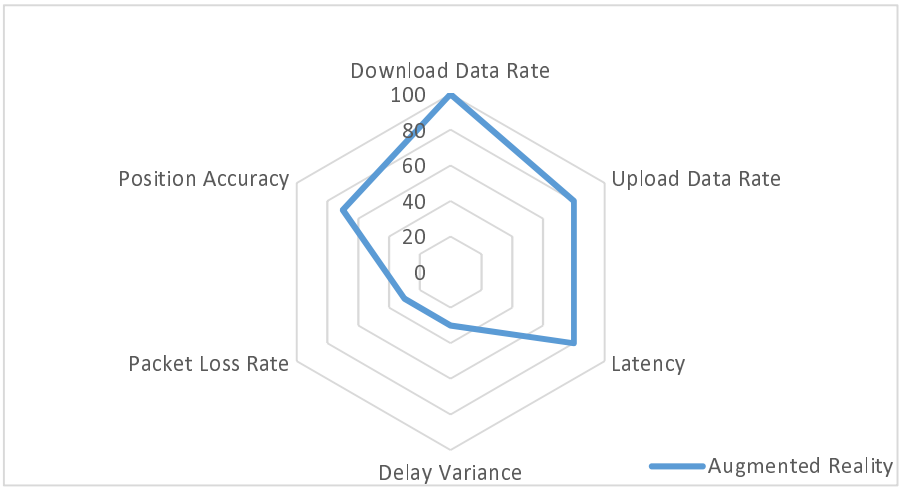

Figure 8: Augmented Reality performance capabilities

The performance criteria for AR scenarios can be identified by the following QoS parameters. Network requirements are identified in clause A.2 of the present document:

- Data rate (DL) (Mbps).
- Data rate (UL) (Mbps).
- Latency/Delay (ms).
- Position Accuracy (%).
- Packet Loss (%).
- Delay Variance (ms).

4.4 User type scenarios

It is anticipated that, 5G will enable use cases across industries for more than just human users. The different user types will place unique performance requirements on the network and invoke distinct traffic patterns. These requirements are dependent on the industry, the type of user and the use case which is being invoked. Traditional networks are optimized for humans and likely not suitable for other user types such as vehicles or industrial IoT. The primary types of users enabled through 5G include:

- Human Users - primary use cases include those identified in relation to increase data rates and lower latency such as those described and supported by enhanced mobile broadband. These applications and technologies are consumed by human users on smart phones or other terminal types such as AR glasses or VR headsets. It is expected that performance criteria are reflective of those identified in clauses 4.1 to 4.3:

  - implies higher data rates, lower latencies and coverage requirements from the entire country to small areas. The traffic requirements are identified in clause A.1 of the present document;

  - implies scenarios and technologies that require high data rates as identified through selected enhanced mobile broadband scenario requirements identified in clause A.2 of the present document is a reflection of use cases, technology and services related to Smart Cities, Smart Home, Connected Energy, and Agriculture amongst others. One of the performance criteria for these type of users include the high availability of IoT traffic, which involves different deployment areas, different device speeds and densities and require a high availability communication service to transfer a low data rate uplink data stream from one or several devices to an application. The requirements identified in ETSI TS 122 261 [i.2], clause 7.5.2 are applicable although not examined in detail within the scope of the present document.

- Automotive - including services and industries in the area of connected car, autonomous driving, vehicle-to-X and other types of use cases in the automotive area. Vehicular scenarios require very low latency and very high reliability. Requirements for V2X in particular are defined in 3GPP TS 26.186 [i.14] and consider scenarios in relation to interworking vehicles platooning, advanced driving, extended sensors, remote driving and vehicle quality of service support. Data rate requirements are not demanding but there is a higher requirement for upload data rates than download in particular for remote driving.
- Industry 4.0 - including services such as those created to support verticals in industries such as manufacturing and private networks. These include services in the areas of motion control, discrete automation, process automation and remote control to identify just some of the scenarios specific to this type of users.

The user type scenarios are represented here to illustrate the types of performance criteria, applicable for different users of the same network. Typically, the QoS parameters can be identified similar to those examined as part of traffic and usage scenarios and require test scenarios and QoS definitions that can accurately measure high data rates, very low latencies, latency variation, information loss, connectivity and coverage.

5 QoS Parameters

5.1 Technical QoS parameters

5.1.1 Overview

Having identified the performance criteria, it is necessary to consider the QoS parameters to measure those requirements.

The aim of the technical parameters is to detail the QoS parameters that can measure and quantify the network capabilities to meet the performance criteria. By measuring these QoS parameters, the traffic, usage and user type scenarios which the network is capable of supporting can be identified.

5.1.2 Coverage

Coverage can be examined from a number of aspects considering the many features available in 5G new radio including, massive MIMO, beamforming, new spectrum availability incorporating millimetre waves and the impact of network slicing to intelligently allocate portions of the spectrum based on specific needs of the application or device. These and other advancements add significant complexity to identify and test the impact of coverage on quality of service. Therefore, the aim is to analyse coverage from the point of view of the impact on throughput/data rates and delay. This requires examination of the impact of coverage level and coverage quality.

Considering the relationship between the downlink throughput and coverage it can be surmised that throughput is directly related to the SINR reported by user equipment indicating that the downlink CSI SINR is a suitable measure for evaluating coverage and interference. Additionally the uplink channel quality of a UE can be indicated by the SINR but this cannot be tested on the UE side and needs to be calculated through relationships to other testable indicators such as the downlink RSRP and the transmit power of the UE and the 5G radio node. Looking at analysis of the relationship between the air interface delay and the CSI RSCP/SINR, the impact on RTT is minimal if the RSRP is greater than -110 dBm and the RTT will deteriorate if the SINR is smaller than 1 dB [i.13].

These measurements need to be taken in sufficient geographical areas to measure the network coverage as previously identified. A typical coverage measurement will measure the land area where the number of samples meets or is greater than the identified New Radio RSRP signal in relation to the total area covered.

5.1.3 Transport

5.1.3.1 Overview

The applications and scenarios described in clause 4.3, identify performance criteria for the more popular technologies and use cases enabled through 5G, initially at least. At the time of publication of the present document, services and applications that require a certain throughput, cope with temporary low data rates or interruptions due to data buffering.

These new scenarios more and more require real time interaction. Consideration needs to be given to the transport mechanisms which are used to properly examine the QoS parameters, and a method to accurately test them and quantify the network ability to support these 5G enabled usage scenarios.

5.1.3.2 Real time considerations

Traditionally TCP, through its guaranteed delivery mechanisms, is the primary method for supporting packet data transport in mobile networks. While this brings reliability, it also brings an overhead in terms of connection establishment, reliability and flow control. In the past, many applications including streaming applications utilize TCP for video streaming [i.3], employing buffering as a means to manage times when the data rate fluctuates. Now UDP, which is long recognized as being suited for applications that require speed and efficiency is seeing increased use in the many real time scenarios associated with interactive technologies such as VR, AR and gaming as referenced in clause 4. Not only is UDP the main protocol for real-time applications, it is also the protocol that comes closest to the physical layer, and it avoids any additional, uncontrollable traffic caused by acknowledgements and retransmissions.

Many real time streaming applications are adopting QUIC [i.15], as the transport protocol. QUIC is a UDP based transport protocol, where the transport is secured on a higher layer with integrity protected header and encrypted payload that operates more efficiently than TCP that can easily be implemented in the application. QUIC is designed to reduce latency, especially for mobile users and can help to improve user interaction responsiveness due to better protocol efficiency with reduced overheads and minimized response traffic. QUIC also minimizes back-traffic caused by per-packet acknowledgements as in TCP.

Today and in the near future, QUIC will serve the delivery of web-content as video information and also web-pages and other content. Interactive applications as video chat or others are still not based on QUIC, rather on unreliable plain RTP/UDP because the applied securing mechanism in QUIC is also based on re-transmission in case of corruption or loss. Today's network latencies do not allow making use of re-transmitted packets for real-time interaction because of their increased delay. However, along with very low latency transmission in upcoming 5G standalone networks, QUIC and its re-transmission scheme may also become usable for real-time interactive applications.

5.1.3.3 Latency & Interactivity

5.1.3.3.1 Overview

An important aspect of data transmission performance of networks, are data transfer times and the resulting reactivity in interactive scenarios. Latency and reactivity are becoming even more essential for new interactive and real-time applications such as Augmented Reality but also in Industry 4.0 or automotive use.

Latency and the resulting reactivity need to be measured in a scenario that reflects the application and use-case to be evaluated. This requires a data transfer profile (traffic pattern) that reflects the application, so that the relevant latency and reactivity can be measured.

A basis for this method is IETF's TWAMP testing protocol [i.17] and [i.21], described in detail in clause 6.2.3 of the present document. In this approach, a scalable UDP packet stream is send from and reflected by a far-end server back to the measurement client, e.g. from a smartphone or modem device to a server in the network or a second device.

The described method to derive latency and interactivity is defined in Recommendation ITU-T G.1051 [i.23].

5.1.3.3.2 Per packet two-way latency

The realization of the described two-way latency measurement method allows the determination of the latency of each individual sent and received UDP packet. As a result of the measurement, the vector *D(i)*, where *D* is the latency of an individual packet *i* for a measurement interval, is available for detailed analysis.

Statistical aggregation metrics of *D(i)* quantiles are recommended:

- Delay D(i) 50^{th} percentile (median).
- Delay D(i) 10^{th} percentile (approximation for the shortest reachable latencies in practice).

An arithmetic average as a statistical mean for characterization of the latency in a measurement interval cannot be recommended, since individual extreme latencies dominate this average. Measured packet latencies follow a so-called heavy-tailed distribution with reduced meaning of arithmetic averages.

5.1.3.3.3 Packet delay variation

In line with Recommendation ITU-T Y.1540 [i.16] and IETF RFC 5481 [i.20], the packet latency of an individual packet is rated to the minimal packet latency and is defined as:

$$PDV(i) = D(i) - D(min) \text{ where } PDV(i) \in \mathbb{R}+$$

where *D(i)* is the individual latency of one packet and *D(min)* is the minimum individual latency of all packets in the measurement interval.

The packet delay variation (PDV) values can only be zero or positive, and quantiles of the PDV distribution are direct indications of delay variation. This vector *PDV(i)* is used to calculate the:

- PDV 50th percentile (median).
- PDV 99,9th percentile ((approximation for the largest reachable delay variation in practice).

The PDV is a relative measure with respect to the packet with the shortest latency. This enables the provision of the PDV per direction, as no time synchronization is needed for relative measures. Based on the timestamps of sending at client side and receiving at server side, the uplink one-way PDV can be computed. Likewise, based on the timestamps of sending at server side and receiving at client side the downlink one-way PDV can be derived. Consequently, PDV is available as:

- PDV_{CS} one-way (client to server).
- PDV_{SC} one-way (server to client).

In principle and if needed, also the resulting two-way PDV (client to server and return) can be obtained from the available timestamps.

In addition to PDV also the interpacket delay variation (IPDV) can be computed. In line with IETF RFC 5481 [i.20], the inter-packet delay of an individual packet is rated to the delay of the previous packet and is defined as:

$$IPDV(i) = D(i) – D(i-1) \text{ where } i \in N$$

where D(i) is the individual delay of one packet.

IPDV values can be both negative and positive, and percentiles of the IPDV distribution are direct indications of delay variation.

This vector IPDV(i) is used to calculate:

- the $IPDV_{>0}$ 34,1th percentile (standard deviation under assumption of normal distribution);
- the $IPDV_{>0}$ 99,9th percentile (approx. maximum).

Here $IPDV_{>0}$ denotes the vector of all positive values of IPDV.

The IPDV is a relative measure with respect to the previous packet. This enables the provision of the IPDV per direction, as no time synchronization is needed for relative measures. Based on the timestamps of sending at client side and receiving at server side, the uplink one-way IPDV can be computed. Likewise, based on the timestamps of sending at server side and receiving at client side the downlink one-way IPDV can be derived. Consequently, IPDV is available as:

- $IPDV_{CS}$ one-way (client to server).
- $IPDV_{SC}$ one-way (server to client).

In principle and if needed, also the resulting two-way IPDV (client to server and return) can be obtained from the available timestamps.

5.1.3.3.4 Disqualified packets

To measure disqualified packets, all packets, which would not be useable for a real and running application are counted. This covers more than actually lost packets, in particular:

- Not sent packets: Packets that could not leave the client device due to uplink congestion and being discarded by the device kernel after timeout.
- Lost packets: Packets that were lost during transmission or could not leave the reflecting server due to downlink congestion and being discarded by the server kernel after timeout.
- Erroneous packets: Packets that were corrupted after arriving back at the client device (see note).
- Discarded packets: Packets that were received back after a pre-defined time-out at the client device. This time-out, also called delay budget, is specified and defined according to the maximum acceptable latency for the target application.

From an application's point of view, there is no differentiation needed between the individual causes of not considering a packet as received. An overall indicator as the Ratio of disqualified packets P_{DQ} (simplified: Packet Loss Ratio) is seen as sufficient at application level.

$$P_{DQ} = \text{number of disqualified packets/number of all packets sent by the application}$$

For a more detailed analysis and trouble-shooting, the individual reasons for disqualifying packets can be reported too.

NOTE: Usually the IP kernel of the operating system discards corrupted packets below the IP interface level. When tracing packets at IP level, corrupted packets are seen as lost.

5.1.3.3.5 Interactivity

Transport QoS parameters such as latency, delay variation and disqualified packets provide performance measurements to evaluate the network capabilities. It is necessary to consider how to relate those measurements to services, use cases and applications to determine the level of interactivity supported. This requires the definition of more complex QoS parameters that will evaluate the network capability in terms of interaction scoring.

The basic concept of scoring interactivity is that latency and the amount of disqualified packets determine the perceived interactivity by a user. It depends on the target application and the user expectation, as to the influence on the perceived interactivity. Two-way packet latency gives information about how fast a response to an action originated at the client user devices is received back. In addition, disqualified packets are missing information for the user's application. Whether they can be interpolated by the application or just lead to temporary distortions or pausing while using the application depends on the application's implementation. An implementation of an approach to score interactivity is demonstrated in clause B.2.1 of the present document.

To receive results for latency and ratio of disqualified packets, the packet stream - especially in data-rate and traffic pattern - should reflect the targeted application in use, either through the capture of application measurements or application emulation.

Consequently, there will not be one single prediction model for interactivity, rather individual ones for different application types.

5.1.3.4 User perceived peak throughput & data rates

To understand the capabilities of data transfer rates and the maximum data capacity that can be supported by a network, it is necessary to assess the peak user perceived throughput measured as the maximum capacity for high speed internet services. Expected 5G data speeds, coupled with content moving closer to the edge, lower latencies and new technologies underscores the need to be able to measure high speed internet access at Gigabit per second rates. Current measurement approaches using TCP have limitations such as sensitivity to packet loss, round-trip time, and flow control overheads. It is necessary to remove the gap between the current measurement schemes and the accurate measurement of high-quality capacity. As discussed, low latency applications are increasingly relying on UDP and protocols like QUIC which encrypt activity above the transport layer.

The gigabit access services delivered by many service providers today have outstripped the ability of ad-hoc, TCP-based methods to measure their performance. The UDP-based IP capacity metric and related measurement methods specified in forums and standard bodies can provide a much more accurate and consistent understanding of network performance.

The maximum IP layer capacity is defined as a UDP-based measurement which measures the maximum number of IP-layer bits (including header and data fields) that can be transmitted from the source host and correctly received by the destination host during one contiguous sub-interval [i.9], [i.16]. It is a measure that represents the user experienced peak throughput rates for applications that utilize UDP as the communication protocol. The test scenario to measure the maximum IP layer capacity is examined in clause 6.

Secure data transfer through HTTPS and HTML 2.0 is the primary method for data transfer across the internet today. It is accepted that latency can increase with HTTPS because of the initial TLS handshake, which requires two extra roundtrips before the connection is established, compared to one through an unencrypted HTTP port. However, HTTP 2.0 advantages through multiplexing, header compression, server push and round-trip time improvements ensure HTTPS adoption. The effective data rate or throughput rate is limited at the top end by the IP layer capacity with the effective rate dependant on the application and the loading conditions. Therefore, while important to measure the capacity, it is also necessary to measure the effective rate for data transfer over HTTPS.

Table 1

QoS Parameter	Maximum IP-Layer Capacity	Data Rate/Throughput
Definition	For a given population of interest, the maximum IP-layer capacity during time interval [*t*, *t* + Δ*t*] is: $Maximum\_C(t, \Delta t) = \frac{\max_{[t,\Delta t]}\left(n_0(dt_n, dt_{n+1})\right)}{dt}$ where *dt* is the duration of a sub-interval. See the reference for further details. The definition is the same regardless of direction of transmission.	Download - the rate at which data can be transferred to a user's device (such as streaming video or playing an interactive VR game). Success Ratio - the ratio of successfully completed downloads to all started downloads. Rate - Sum of all DL transferred bits over the active transfer period. Upload - the rate at which data can be transferred from a user's device (such as streaming real time content to social media sites). Success Ratio - the ratio of successfully completed uploads to all started uploads. Rate - Sum of all UL transferred bits over the active transfer period.
Reference	[i.9], clause 4.5.	[i.1], clause 6.8.7; [i.9], clause 4.2.

5.2 Service QoS parameters

5.2.1 Overview

The measured network capabilities determine the usage scenarios, technologies and services that the network can support, be that high bandwidth or low latency services. To determine the level of that impact, parameters to measure the service quality are examined through parameters which specifically measure 5G supported services and 5G enabled technologies. These QoS parameters are based on existing technical specifications where available. Of course, percentiles and rates of these QoS parameters are also typically measured in network benchmarking as defined in ETSI TR 103 559 [i.19].

5.2.2 Telephony service

Voice services continue to be an important service in 5G, expected to be characterized by higher accessibility success rates, less cut offs and better quality and is represented by existing QoS parameters. The performance criteria and subsequent QoS parameters for voice services will not change significantly and are highlighted here for completeness.

Table 2

QoS Parameter	Definition	Reference
Call Setup Success Ratio	The success ratio of the voice service independent of access technology is defined in ETSI by the Telephony non-Accessibility. The telephony service non-accessibility denotes the probability that the end-user cannot access the mobile telephony.	[i.1], clause 6.6.1
Call Drop Ratio	Reliability measured as the call drop or cut off rate denoted by the probability that a successful call attempt is ended by a cause other than the intentional termination by either party.	[i.1], clause 6.6.5
Call Setup Time	The call setup time starts with the initiation of the call and ends when the alerting of the called side is indicated. Alternatively, the time where the acceptance or successful setup of the call is signalled to the user can be used as the end trigger.	[i.1], clause 6.6.2
Speech Quality	The telephony speech quality is an indicator representing the quantification of the end-to-end speech transmission quality of the mobile telephony service. This parameter computes the speech quality on a sample basis using MOS-LQO scales.	[i.1], clause 6.6.4

5.2.3 Data service QoS parameters

5.2.3.1 Overview

The network data capability supports existing services and technologies including web browsing and video as well as enabling and enhancing services and technologies such as virtual reality and cloud gaming. These have network performance criteria in relation to increased data rates, reduced latency and higher reliability.

5.2.3.2 Web browsing QoS and HTTPs

HTTP browsing tests the quality of service in relation to website access. These websites can be static type test pages such as the Kepler page or dynamic to represent typical or popular web pages accessed by users on a regular basis. Up to now, HTTP web browsing is measured by success ratios which are defined based on web sites browsed with a return of HTTP status code 200 compared to the total websites accessed as defined in ETSI TS 102 250-2 [i.1], clause 6.8.1 and ETSI TS 103 222-2 [i.9]. Additionally, the time to access a web page is defined as the time from request by a URL until the connection to the content server is established and/or the first byte of payload is received and this includes DNS-resolution and potential re-directions.

Web pages are almost completely delivered over HTTPS now, where there is no longer an OK status. Usually, the browser indicates 'page complete' but 'completion' is dependent on the size of the page, the complexity of the page with regard to images, live information, etc. At the time of publication of the present document, the ETSI Working Group STQ Mobile was working on a new report which aimed to better measure web site browsing to reflect the changes in web site delivery over the last number of years where success criterion based on a standard download size or time will be incorporated.

It is expected that web browsing access times which measure the QoS experienced at the user terminal will be measured by the time from requesting a web-page through a URL until a defined amount of content data is delivered. The defined amount of data is the key definition here and on-going work is looking at determining the patterns and models that can represent sufficient data for displaying the main content of the webpage. This time considers apart from DNS-resolution and contacting the server also the ramp-up of delivery speed and potential opening of connections to additional resources. It is expected that a QoS parameter in this form will explicitly measure Time to X bytes as an access time representation.

Moving away from page complete also makes the measurements less dependent on which websites are chosen for the testing.

5.2.4 Enhanced UHD video QoS

It is expected that higher video definition will become increasingly common in the form of ultra-high definition such as 4K. To date the primary video streaming use case have been related to streaming in the download direction. Already there is a large amount of interest in uploads but with 5G it is expected that this will increase especially in relation to social media. The QoS parameters for video streaming will not change substantially from those already defined for existing streaming services. Video streaming and real time video will place demands on the network in terms of data rates and latency.

Table 3

QoS Parameter	Definition	Reference
Video Streaming Success Ratio	The video streaming success ratio is the end-to-end success ratio of the requested video stream. It starts with the request of the video and ends with the end of the playout. Typically, it is supported by a timeout or cut off time.	[i.1] and [i.3]
Video Access Time	The video streaming access time/initial buffer delay is the time from stream request to the display of the first picture and start of playout i.e. the time it took for the video to start displaying. It includes the video preparation time and the pre-playout buffering time or in the case of real time video, it is the time to access the streaming video feed.	[i.1] and [i.3]
Video Freezing Ratio	The proportion of video freezes or stalling in a streaming session proportional to the display duration.	[i.3], clause 4.5.4
Video Resolution	The resolution is determined by the video content and device information in terms of width and height and in screen pixels. Adaptive streaming mechanisms will ensure that data is delivered at the resolution supported by the delivery components and may change during playout. Therefore, the average resolution should be considered and provide context to the expectation i.e. HD, UHD-1, UHD-2.	[i.3]
Video/Visual Quality	An integrative, comprehensive measure for the quality of the displayed video that combines the impact of freezing and compression including resolution and other spatial and temporal artefacts.	[i.1], [i.3] and [i.18]

The quality of new services and technologies enabled by eMBB, can be identified, in the main, through QoS parameters similar to those of streaming services such as video. There are specific QoS considerations depending on the service usage scenario. In particular, there are QoS parameters which are necessary to measure the network capability to support interactive applications that may not be necessary or suitable for the more graphic focused options. Examples considered here are Virtual Reality and cloud gaming.

5.2.5 Virtual Reality QoS

VR inherits parameters from the streaming scenario but there are VR performance criteria which necessitate specific QoS parameters. This is the case particularly for interactive VR such as VR gaming, which has specific requirements in relation to the user interaction that is dependent on the network latency capability. As identified in ETSI TR 126 929 [i.5], there are a number of VR metrics which are relevant in terms of media transmission, including throughput, buffer levels, presentation delay, playback operations and measurement of interaction. In addition, VR quality of experience is examined in Recommendation ITU-T G.QOE-VR [i.7], with analysis of system influencing factors in including network/transmission related factors, media related and task context which form input to this analysis. Key factors in service presentation measure the continuous and smooth sensory experience of the VR service, where poor user experience is identified by tiling artefacts and stalling in the service quality.

The following QoS parameters relate to Cloud VR Video and Cloud VR Interactive usage scenario performance criteria. VR will place substantial demands on the network in terms of data rates, latency and latency variation.

Table 4

QoS Parameter	VR Category	Definition	Reference
VR Streaming Success Rate	Cloud VR Video Cloud VR Interactive	The VR streaming success ratio is the end-to-end success ratio of the requested VR stream. Success Rate [%] = (VR playout success/VR attempt) × 100, where (playback success = successfully buffered) See note 1.	[i.5]
VR Access Delay	Cloud VR Video	The VR access time/initial buffer delay is the time from VR request to the display of the first VR picture. Access Time [s] = (t start of VR playout - t request of the content) See note 2.	[i.5], clause 9 [i.7]
VR Freeze Rate	Cloud VR Video Cloud VR Interactive	The VR freeze rate is the accumulated video freezing duration in relation to the VR playout duration. Freeze Rate = (VR playouts with freezing /VR playout success) × 100	[i.5], [i.4], [i.7]
VR Tiling Artifact/ Mosaic Rate	Cloud VR Interactive	For interactive VR such as VR gaming, tiling artifacts/mosaic impacts on the smoothness and continuity of the VR playout where mosaic in some areas of the image. During VR gaming, if some video frame information (some block information in the video frame) is lost, the image can be displayed, but mosaics can occur in some areas. The VR mosaic rate is the accumulated video mosaic duration in relation to the VR playout duration. Mosaic rate = (VR playouts with artefact error/VR playout success) × 100 See note 3.	[i.5], [i.7]
VR Resolution	Cloud VR Video (360)	To achieve better VR video quality, a 4K or higher resolution is required, because the VR has a 360 panoramic display function, and the single-eye resolution of the VR panoramic display determines the VR image quality. See note 4.	[i.4], clause 9 [i.5], clause 10
Visual VR Quality		An integrative, comprehensive measure for the quality of the displayed VR visual information that combines the impact of freezing, and compression including resolution and other spatial and temporal artefacts.	
NOTE 1: VR 360 similar QoS measures to video steaming including stop reason from event PlayList. NOTE 2: For VR Interactive such as gaming, the delay is more dependent on the gaming server and device and less related to network. NOTE 3: Visual impact of presentation delay and packet loss. NOTE 4: VR 360 supported through similar QoS measures to video steaming to measure delivered solution meets requirements.			

The quality of service of VR real-time interactive scenarios such as gaming, is primarily impacted by the action response delay. This is designated as the motion to photon delay which is measured by the re-rendering of images within the field of view of a VR device. This requires cloud servers to transmit the updated media stream over the network in a timeframe that ensures the view is displayed in a natural way for the user.

Calculating the motion to photon QoS parameter as the primary measurement of presentation delay requires measurement of the time taken for each of the primary components. That is the cloud server response time, the network transmission time and the device decoding and buffering time:

- Cloud server delay: the cloud processing delay includes the time taken for the VR cloud server to obtain the action /motion, the rendering delay, the encoding delay and the delay to prepare and send the response.

- Network delay: this includes the uplink transmission delay and downlink transmission delay. The sum of the two delays is close to the network RTT which needs to be measured in an accurate manner as per test scenario in clause 6.2.3 of the present document.

- Device processing delay: which includes the buffering delay and device decoding delay.

It can be surmised that the average value of the total cloud and device-processing delay be considered as a constant with the primary delay attributed to the uplink transmission delay, the downlink transmission delay and the buffering delay at the device. The buffering delay is dependent on the data transmission throughput i.e. the delay is a function of the data frame size server encoding for each game image frame in relation to the data rate.

Therefore, after the server and device delays are determined, the motion to photon is directly limited by the network latency and network bandwidth assuming the game bit rate and frame rate are fixed. From the QoS parameter definition is can be see that, optimizing the RTT and throughput is the key to reducing the MTP latency.

Table 5

QoS Parameter		Definition	Reference
VR Motion to Photon Delay	Cloud VR Interactive	The VR Motion to Photon delay refers to the response duration of the video and audio after a user performs an action during VR experience. Motion to Photon [ms] = (t start of action - t action response). See notes 1 and 2.	[i.4], [i.5], [i.7] [i.13]
NOTE 1: This can be measured as: MTP [ms] = (cloud server + device, constant delay) + network delay + (bit rate/frame rate)/network data rate.			
NOTE 2: While not exactly the same, MTP is most practically the main reason for presentation delay.			

5.2.6 Cloud gaming QoS

Cloud gaming similar to VR and video is measured by QoS parameters that evaluate streaming quality. The capability of the network to support performance criteria for two primary game categories, one highly graphical and the other highly interactive, is reflected in the quality of the service delivered. Cloud gaming will place substantial demands on the network in terms of data rates, latency, latency variation and packet loss.

Cloud gaming response delay is a key measurement factor and is made up of:

- the command data collection delay i.e. collecting the user input on the client;
- the uplink transmission delay of the user command data;
- receiving, processing and rendering the picture in response to the command on the gaming server in the cloud;
- graphic preparation delay at the cloud service in terms of graphic capturing and coding delays;
- the downlink transmission delay of the coded graphics to the client;
- the graphic decoding delay at the client;
- client rendering delay, including the delay in waiting for the video buffer after decoding and colour processing;
- delay in displaying to the user.

Table 6

QoS Parameter	Definition
Cloud Game Streaming Success Rate	The cloud game streaming success ratio is the end - to - end success ratio of the requested cloud game.
Cloud Gaming Freeze Rate	The cloud gaming freeze rate is the accumulated game freezing duration in relation to the playout duration.
Cloud Gaming Tiling Artefact/Mosaic Rate	For interactive gaming, tiling artefacts/mosaic impacts on the smoothness and continuity of game play. During gaming, if some information is lost, the image can be displayed, but mosaics can occur in some areas.
Cloud Gaming Response Delay	The cloud gaming operation delay refers to the response duration of the game after a user performs an action.
Visual reproduction quality	An integrative, comprehensive measure for the quality of the displayed visual information that combines the impact of freezing, and compression including resolution and other spatial and temporal artefacts.
NOTE: Cloud Gaming Access Time is not deemed relevant as the delay is more dependent on the gaming server and the user device and less related to the network.	

6 Test scenarios

6.1 Overview

Defining and executing the test scenarios that will measure the network capabilities is a critical part in the deployment of 5G. Accurate network assessment will ensure there is confidence in the quality levels of services and applications that the network can support. Extremely high data rates and very low latency are expected from 5G, with many applications using UDP as the transport mechanisms requires specific ways to test and measure the network performance. These methods provide testing techniques that reflect the demands of the applications and services that will utilize the network. In addition, test scenarios to measure the impact of the network capability on the enabled services will be examined. In the main that will consist of existing measurement methods for existing services but also examine the usage scenarios identified for high data rate enabled services.

6.2 Executing network test scenario

6.2.1 Overview

The measurements streams need to reflect one of two scenarios to accurately measure the capacity and throughput date rates including the packets that are corrupt or lost along with the need to measure down to very low latencies. These scenarios are:

- Very efficient transmission of IP/UDP packets to ensure that maximum capacity is actually assessed.
- Real time application transport mechanisms through UDP.

6.2.2 Measuring maximum user perceived throughput and data rates

Current measurement methods are primarily TCP based, reflecting the protocol attributes such as flow control and retransmission, reacting conservatively to loss and round-trip delay. This leads to underestimating of the actual data rate throughput and capacity levels achievable and TCP not reflective of the many applications that now utilize UDP. Measuring the maximum user perceived throughput as the IP-Layer Capacity on an access link should use the same transport protocols as used by these applications. To this end, effort applied towards industry harmonisation on the Maximum IP-Layer Capacity Metric and Methods are being realized across industry standard organizations.

Recommendation ITU-T Y.1540 [i.16] defines a set of IP-layer metrics and methods of measurement with the maximum IP-layer capacity of interest here. The definition for the end-to-end case, is based on the total packets sent from source to destination using UDP as the transport mechanism where the measurement points represent the source and destination locations with respect to a user's expected position and the Internet access link being measured. Rather than repeat the method of measurement from [i.16], a summary of the main steps will be described with the condition that any implementation of this approach will follow the full method in that reference. Therefore the primary steps are that:

- The sender arranges to send and receive the stream of IP packets using UDP transport-layer with key parameters which are defined in the reference but include packet type details, sending rate and other characteristics.
- During a test, the sending rate is varied in accordance with a specified search algorithm, search goal with defined measured metrics and thresholds, duration and sub interval rules and limits.
- Result collection, validation, and use of consistent criteria and reporting specification follows.

The primary output from the test will be to report the maximum IP-layer Capacity, the trial IP packet loss ratio, and other metrics when available. The measurement system may also report UDP capacity in terms of UDP payload bits delivered, because this is the capacity available to user applications after IP and UDP headers have been removed. Round-trip Latency and variation, one-way delay variation, and packet reordering are also easily measurable using the same UDP stream.

While capacity measurements determine the maximum transfer capability of the network, the data rate measurements explain the practical measure of actual packet delivery, albeit affected by factors beyond the service provider's control such as user equipment and software configuration, browser version and configuration, and the details of the path between the user and their desired content. As it is a measure of how much data can be sent and received at a time, it is best measured through the application of specific service data requirements in typical network conditions. This means the transmission of data packets which reflect typical user scenarios in the test environment. This can be achieved through generating traffic patterns that reflect testing requirements such as quick test measurements that send and receive a basic package pattern or through more focused testing with different packet sizes, sending rates and timeouts. Testing is to measure data rates in uplink and downlink directions to represent typical activities is necessary. This will give a representative view of the network based on the defined traffic pattern but it is dependent on the traffic patterns accurately reflecting user activity.

So, while the Maximum IP-Layer Capacity measurement approach measures the network capability, it is representative of a technological limit and prevailing conditions at the time of measurement, and results are expected to align with the service provider's description of Internet access. This testing is complemented by measurement of the data rates captured through the actual use of data services through typical user actions. For example, data rate measurements over HTTP, reflect the QoS of user actions through web browsing. The relevant test scenarios in relation to access, transfer and tear down are defined as part of ETSI TS 102 250-3 [i.10], clause 9.3 and equally apply to 5G. While, the reference refers to HTTP, the expectation is that measurements over HTTPS need to be evaluated here. Typical test scenarios measuring data rates that reflect user activities in a typically loaded network include:

- Fixed duration and fixed size file upload and download.
- Web Browsing of popular, static and dynamic sites.

6.2.3 The TWAMP method to obtain two way latency

Applications technologies and services are increasingly providing real time services or require very short delay and high levels of interconnectivity. This is evident from earlier descriptions in areas such as interactive VR applications, cloud gaming applications and others. A test scenario and method in which to measure these very low latencies as well as the ability to maintain connectivity is necessary. As discussed previously the importance of UDP in real time scenarios necessitates the measurement of network latency capability over the same transport mechanism.

To date ping tests are commonly used as a means to test round trip latency. Ping tests use a dedicated protocol (ICMP) but it is not reflective of real applications, load patterns or data flow. A number of measurement methodologies can be used, based on UDP where the two-way latency of each individual packet can be obtained. The preferred method to obtain two-way latency is the IETF Two Way Active Measurement Protocol (TWAMP) methodology and protocol specified in IETF RFC 5357 [i.17], supported by additional features defined in IETF RFC 6038 [i.21], where the responding host returns some of the command octets or padding octets to the sender, and provides an optional sender packet format that ensures equal test packet sizes are used in both directions. TWAMP is based on a UDP packet stream of packets of pre-defined size and frequency and depending on the vendor - supported by infrastructure components such as routers and IP-gateways.

The TWAMP measurement architecture is usually comprised of two hosts with specific roles, and this allows for some protocol simplifications, making it attractive solution. The basic architecture is as in Figure 9.

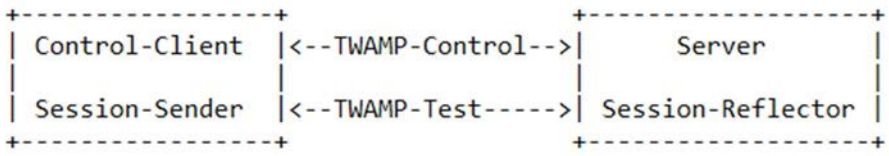

Figure 9

TWAMP consists of two inter-related protocols: TWAMP-Control and TWAMP-Test. TWAMP-Control is used to initiate, start, and stop test sessions, whereas TWAMP-Test is used to exchange test packets between two TWAMP entities. One entity being the Session-Sender and the other being the Session-Reflector. The Session-Sender is the entity that sends the packets and also collects and records the necessary information provided from the packets received from the Session-Reflector. TWAMP requires the Session-Reflector to transmit a packet to the Session-Sender in response to each packet it receives. The received reflected packet can be assigned, to a sent packet, by an ID. The difference of the sending and receiving time stamps are reported as two-way latency.

The amount of data to transmit and the resulting data rate can be specified by packet size and sending frequency. This allows an emulation of data stream characteristics as produced by real applications. An important decision is in the location of the Session-Reflector and its relationship to actual enterprise deployments for applications such as VR or cloud gaming. There are options to consider with regard to testing at the network edge, in the cloud or on premise. It needs to be acknowledged that the further the reflector is from the network edge, the more the latency is dominated by the content delivery network.

Therefore, in the case where the requirement is to solely measure the network capability, test scenario setup requires that the Session-Reflector is placed as close as possible to the network edge. However, where the requirement is to measure the network capability closer to a VR or gaming application scenario, it is probably more suitable to place the Session-Reflector in the cloud or in a private network.

In addition to testing and measuring the latency, the delay variation measured as jitter and packet loss are tested and measured as part of the TWAMP deployment:

- The delay and jitter are measured based on timestamps where the sender sends a packet with sending timestamp T0, and the reflector replies with a receiving timestamp T1 and a responding timestamp T2. When the sender receives the response, it records a receiving timestamp T3. The delay and jitter are calculated based on these timestamps.

- Packet loss is calculated based on serial numbers carried in the packets. The sender packet contains a serial number, which the reflector uses to respond with the same serial number. Each packet serial number is increased by the sender and the reflector. The packet loss is calculated based on the difference between the sent and received packets.

6.3 Executing service test scenarios

6.3.1 Overview

Having measured the network capabilities with regards to latencies, data rates and reliability, it is necessary to examine how these capabilities are reflected in actual use of a service or application. Typically, this will present quality impairment in terms of times taken to perform an action or in the delivered quality which can be represented visually or in the interaction experience. The test scenarios to assess this impairment and provide the data to populate the previously identified QoS parameters are considered here.

6.3.2 Scenario identification

6.3.2.1 Overview

Test scenarios are defined through a reusable, repeatable and standard manner as described in ETSI TS 102 250-3 [i.10]. The definition of service QoS is in general well defined for existing services and technologies from voice to video and referenced where appropriate. For 5G enabled services and technologies, applying this method [i.10], identifies test scenarios, and is applied here for examples in VR and cloud gaming. The method is composed of a number of stages:

- Setup and control [i.10], clause 6.1 where, measurements should be conducted in a way that user behaviour is realistically modelled. Parameters and settings which have substantial influence on results need to be under control of the measurement equipment.
- Phase classification composed of:
  - Service Access consisting of the steps leading to the technical ability to perform content transport.
  - Service Usage which consists of the actual content transfer.
- Result classification which can be:
  - Failed, where the phase of service usage under test was not reached.
  - Completed, where all content intended to be transferred has been successfully transferred.
  - Dropped/Cut Off, where the service usage was ended before completion.

6.3.2.2 Testing methods

6.3.2.2.1 Overview

The measurement methods of testing considered are through testing using simulated actions or testing with user traffic. Testing using simulated actions, performs analysis based on sending service modelled traffic to a destination through a typical representative application emulating service modelled data pattern. Testing with user traffic performs analysis based on collecting actual user application service quality data either via probing, OSS systems or crowd sourced data.

6.3.2.2.2 Guidelines for testing with real applications

In the use of real user applications for testing, the primary consideration is the data accessibility. Data for existing services such as web browsing [i.9] and video streaming [i.3] will continue to be generated and collected as per existing methods. For new services and applications such as VR or cloud gaming, similar data accessibility is necessary. There is an increasing propensity for applications to utilize proprietary encrypted communication. In that case, engagement through standards or the use of application SDKs offering the exposure of measurements from test devices, collected directly from applications or from application servers will be necessary.

6.3.2.2.3 Guidelines to derive traffic patterns for the emulation of real applications

When creating a traffic pattern, the goal is to derive an archetype traffic pattern that is representative of the target application. This should usually cover the main usage scenarios of a group of similar applications like high-interactive e-Gaming, remote drone control or highly visual applications such as VR Cloud Gaming. Alternatively, the traffic pattern could also be targeted to an individual application and use-case.

First, the traffic created by several representative real applications for all usage scenarios should be analysed. The IP trace should be recorded in the different phases of application usage, e.g. initializing phase, active interaction, passive use, and trailing phase. For all applications, use-cases and usage phases, the traffic should be analysed regarding uplink and downlink bitrate, packet size and frequency.

The next step is the definition of a traffic pattern by segments of different bitrates. The proportion of bitrates and their profile vs. time should be driven by the real use but in a shorter overall duration (e.g. 10 to 15 s, where segment duration can be down to 1 s). The order of the segments should resemble the real time profile, e.g. start with an initializing phase and not a highly interactive phase. In asymmetric traffic scenarios, the uplink and downlink bitrates can differ and have a different proportion for each segment.

Finally, the target bitrates for the segments have to be broken into the parameters packet size and packet sending frequency. The packet size is limited to the range between the minimum defined by the header size of all transport protocols including the TWAMP header and the maximum defined by TWAMP (~65 000 bytes) [i.21]. Within this range, the packet size and frequency should be set to resemble the real application traffic. One additional restriction applies to the packet frequency: it should be identical for uplink and downlink traffic within one segment, because of the round-trip nature of the measurement. Packets are only reflected at server side, not multiplied nor discarded. The application of these emulation guidelines is consolidated by a practical example in clause B.2.2 of the present document.

6.3.2.3 Classification of measurement environment

For interpretation and comparability of test results it is necessary to recognize the test measurement environment to ensure consistency and relate the test results to the performance criteria. Typical environments are detailed in clause 4.1 and include urban macro, rural macro, indoor, dense urban, etc. The measurement environments also highlighted in ETSI TS 102 250-3 [i.10], primarily cover similar environment types. It is expected that full 5G coverage will be verified at test locations both indoor and outdoor prior to the execution of 5G testing. In some locations, test design and planning need to consider limitations imposed by the suitability of environment or service under test in terms of equipment, power and mobility.

6.3.3 Impact of 5G features and application intelligence

It is acknowledged that 5G specific features such as network slicing, context aware networks, dynamic policy control, QoS control and others [i.12], will need consideration in test scenarios. This includes the necessity to consider multi service and multi device test setups to effectively invoke and assess network behaviour dependant on feature deployment and its impact on user options. An example of this is with network slicing, where it may be necessary to test the network capability or a network service by invoking different offerings, possibly through different network slicing options that may support a service and measuring the network capability against different performance criteria.

In addition, ML and AI in applications will mean constant adjustment and adaption of application based on network conditions. This will involve prediction context awareness that may for example adapt resolutions to ensure continuity, requests for additional bandwidth in anticipation of network resource competition or requests for prioritization to meet latency conditions.

These features and technical intelligence imply that test scenarios in relation to services and supported applications need to consider an approach that take these factors into account.

6.3.4 Test scenarios

6.3.4.1 Overview

Test scenarios for existing services will not change significantly and are referenced as transaction definitions and transaction types [i.10] and related service profiles [i.11].

6.3.4.2 Telephony testing

Telephony measurements, including transaction definitions, parameter overviews and result definitions [i.10], clause 7, equally apply to 5G. Additionally Telephony service profiles identifying transaction durations and call windows are suitably identified in ETSI TS 102 250-5 [i.11], clause 4.2.1.

6.3.4.3 Video streaming testing

Video measurements, including transaction definitions, parameter overviews and result definitions are identified in ETSI TS 102 250-3 [i.10], clause 9.6 and equally apply to 5G. It should be noted that the most popular video services today are OTT applications which are outside the control of the network operator. However, either testing through the generation of real time video packet data patterns or testing through the use of actual user applications needs to be considered.

In most cases a video service can be divided into the setup of the context until the media server links are received and the phase where the video data are retrieved and the video playout [i.3]. The three primary measurements collected during video streaming [i.3], are related to failure to start, video freezes and low-quality resolution of the received video.

As discussed earlier, for 5G, it is expected that high resolution video streaming will become the norm in terms of UHD, 4K and 8K content. To reflect this user experience video streaming scenarios should test content with resolutions that reflect these requirements. Additionally, the ability of the test equipment to support such high-resolution content. In particular, the device firmware and the network limits need to be considered.

6.3.4.4 Virtual Reality testing

Virtual Reality as one the primary technologies enabled by 5G is examined as an example of a new service. Significant analysis has been performed especially with regards to QoE parameters and metrics [i.6], including a number of conceptual observation points which would be suitable for specific metric-related data collection. These observation points need to be able to communicate the data to measurement collection systems. It is likely that the optimal method to collect this data is through the VR application and the device. One way to do this is to create a test application built using a VR platform SDK that collects performance information on the device. The other is to align with virtual reality industry players to expose the required application and device information in a standardized way.

Of primary interest are the test scenarios to support measurement of transmission and device impact on VR QoS. These measurements should be collected based on various test scenarios in applications supporting VR Video and VR interactive categories. The test scenarios can be characterized through access to content, content playout, in play interaction through applications that reflect the primary VR categories of VR video such as 360 degree videos and VR interactive such as VR gaming.

The primary measurements to be collected during VR testing, reflect those for other streaming services such as video and measured in terms of:

- failure to start;
- freezes and other quality measurements;
- interactive response delay and impacts.

Fully interactive cloud VR applications are necessary for testing to ensure that the testing is not simply a VR game run locally with an interaction element. This local execution of the application has much lower throughput requirements as the multi-player aspect does not stress the network sufficiently, although latency continues to play an important role. The test scenario needs to ensure that the interaction element e.g. game play is smooth and not impacted by delays, elements are where they are expected to be e.g. game opponents and there is no serious visual quality impacts such as graphical glitches, delays or artefacts observed during testing. Additionally, issued commands are expected to be responsive with non-existent input lag.

6.3.4.5 Cloud gaming testing

A further example of a service expected to benefit from the higher data rates and lower latencies of 5G is cloud gaming. Test scenarios, somewhat similar to video and VR necessitate measurement at the end user device. Measurements reflect a user's ability to access the service and the quality of that service, especially with regards to the visual and interactive component. For cloud gaming the game picture is calculated and rendered on the cloud with the rendered picture transmitted to clients across the network.

Cloud gaming platforms differentiate based on the service they offer, where for example one provider may provide a high-end personal computer option in the cloud from which games can be played on a local terminal, others provide gaming service streamed directly to the terminal. The test scenarios should reflect the different offering as there are different implications for data rates and latencies. Where the cloud gaming is streamed from the cloud, the download will be more stressful to the network. However, the latency is still extremely important due to reactive uplink commands from the device to the game. Additionally, the location of the cloud game servers is an additional important consideration for testing. In some instances, the latency and throughput may indicate a good enough performance to play a game, but may not provide the same quality at a higher resolution and need to be tested against a number of different levels of performance criteria.

As per VR test scenarios, data measurement collection at the user terminal or application using actual user application is the most accurate test scenario for cloud gaming. The test environment could be simulated based on a client-cloud server architecture with appropriate responses and measurements at the test client of the latency and throughput of typical packet data that reflects the types of games being tested. This does require extensive analysis and modelling to ensure the data generated at different phases in a typical cloud game are accurately reflected.

7 Summary

It is expected that 5G will deliver significant benefit to users in terms of experience as well as enabling new use cases and new industries. These benefits will be in terms of much higher data rates, lower latencies, better energy efficiency, great reliability and availability. As 5G is deployed and matures, it is necessary to be able to measure the capability of the network to determine what technologies, applications and use case scenarios the network can support. If an industry or a specific use case requires data rates in the gigabit range or latency down to a few milliseconds, it is necessary to test the network in a manner that can accurately measure the capabilities. The present document examines test scenarios to measure those capabilities whilst also considering, the impact on the primary technologies, applications and services enabled through the higher data rates of eMBB.

Whilst the methods and measurements described can be used independently, the present document first examines the performance criteria for different scenarios including traffic, service and user. This provides guidance to identify performance criteria recommendations for these scenarios. It examines the QoS parameters to effectively evaluate those performance criteria and finally defines the test scenarios that can support the measurement of those QoS parameters. The present document investigates developing methods to measure technical network parameters that test the high data rates and low latencies available through 5G. It considers the impact of those network capabilities on the service, technologies and use case scenarios that 5G will enable through examples in ultra-high definition video, virtual reality and cloud gaming.

Annex A:
Performance recommendations

A.1 Traffic scenario criteria

A.1.1 High data rates and traffic density performance criteria

Examples of high data rate and traffic scenarios are given in [i.2]. The scenarios cover examples as "Airplane connectivity" with a recommended experienced downlink data rate of 15 Mbit/s up to "Indoor hotspots" with an experienced downlink data rate of 1 Gbit/s.

A.1.2 High data rate and low latency performance criteria

Examples of high data rate and low latency performance criteria are given in [i.2] too. The scenarios cover examples as "Cloud/Edge/Split Rendering" or "Gaming or Interactive Data Exchanging" with a recommended end-to-end latency of 5 ms to 10 ms and a high reliability of 99,99 %.

A.2 Service Scenario Criteria

A.2.1 UHD Video performance scenario

Examples of UHD video performance scenarios are given in [i.13]. There are recommended data rates in downstream of 30 Mbit/s for on demand UHD video and 40 Mbit/s for UHD live video. There are also criteria of a round trip time for data packets of < 100 ms and a maximum loss rate of 10^{-3}. There are also examples of real measurements in UHD video services listed in [i.13].

A.2.2 Virtual Reality performance scenarios

The performance scenarios in [i.13] also cover Virtual Reality services, here data rates for different image resolution from Full HD (8 Mbit/s) up to 100 Mbit/s for most demanding UHD high resolution cloud gaming are listed.

A.2.3 Cloud gaming performance criteria

- 720P or 1080P, with frame rate of 30 fps, which is the game resolution that can be provided by mainstream cloud gaming platforms. Whether the game is displayed on a mobile phone screen or a monitor, the experience just exceeds acceptable.
- For 5G cloud gaming, it is recommended that an experience of 1080P@60 fps be given first, corresponding to professional players. In most scenarios (operation response delay 70 ms), a good experience is expected.
- For a large display screen, when the resolution of a game is 2K (2 560 x 1 440), the frame rate is 60 fps, and the operation response delay is 50 ms, user experience is very good.
- When the resolution of a game image is 4K (3 840 x 2 160), the frame rate is 60 fps, and the operation response delay is 50 ms, the experience is excellent.

A.2.4 Augmented Reality performance criteria

In AR the latency requirement for Motion-To-Photon (MTP) latency is even stricter than in VR as visual changes are not only triggered by the motion of the user but also by any change (e.g. lighting or natural object movement) in the surrounding world.

In [i.2] latency performance criteria are listed. In uncritical health and wellness scenarios maximum 100 ms end-to-end latency is recommended, for real time services the recommended latency is 20 ms.

Annex B:
Emulation & Interactivity example

B.1 Definition of test cases

The test case for deriving a prediction of interactivity is mainly defined by the used traffic or bitrate profile including packet size and frequency, and possible model parameters to rate perceived interactivity. As additional parameter definition in a test, the delay budget has to be set. A packet with a two-way latency exceeding this budget is considered as disqualified, means discarded due to late arrival by the emulated application.

Limits for latency of different application classes, grouped by the standardized 5QI value, are given in ETSI TS 123 501 [i.22]. They are preferably to use if the emulated application matches, otherwise best practice assumptions or actual discarding limits of the applications are practicable.

B.2 Application emulation and interactivity model parameters

B.2.1 A generic interactivity model approach

The generic interactivity model approach is defined in Annex A to Recommendation ITU-T G.1051 [i.23]. The basic assumption of modelling perceived interactivity is its monotonous dependency on data latency. The shorter the data transport time, the shorter the response time in an interactive application and the perceived use of the application is considered more interactive.

However, this dependency is not a simple linear function; rather there are saturation areas at both tails of the function, where no change in perception happens anymore even the latency changes.

A possible approximation of this non-linear dependency is a logistic (sigmoid) function.

$$f(t) = 1 - \frac{1}{1 + a\ e^{-b\,t}}$$

where *a* defines the horizontal shift on the *t*-axis and *b* the gradient.

For a parametrization that matches to the value range of positive latencies and a scaling by f_0 to a maximum score value of $f_P(0) = f_{max}$, the following formula is applied:

$$f_P(t, i) = \frac{f_{max}}{f_0}\left(1 - \frac{1}{1 + e^{-\frac{1}{b}(t_i - a)}}\right) \text{ for t > 0 with } f_0 = 1 - \frac{1}{1 + e^{\frac{a}{b}}}$$

where *t* is the packet latency in milliseconds, *i* is the indicator of the packet, *a* defines the shift of $f_P(t, i)$ directly on the *t*-axis in milliseconds and *b* defines the gradient of $f_P(t, i)$, where larger values of *b* make $f_P(t, i)$ less steep

The scaling factor f_0 guarantees the maximum score value at $t = 0$. In case, the upper saturation area of the $f_P(t, i)$ starts in the range of $t > 0$, the f_0 will be close to 1. If $f_0 \rightarrow 1$, the parameter *a* defines the latency value where the perceived interactivity is fallen to almost 50 % of f_{max}.

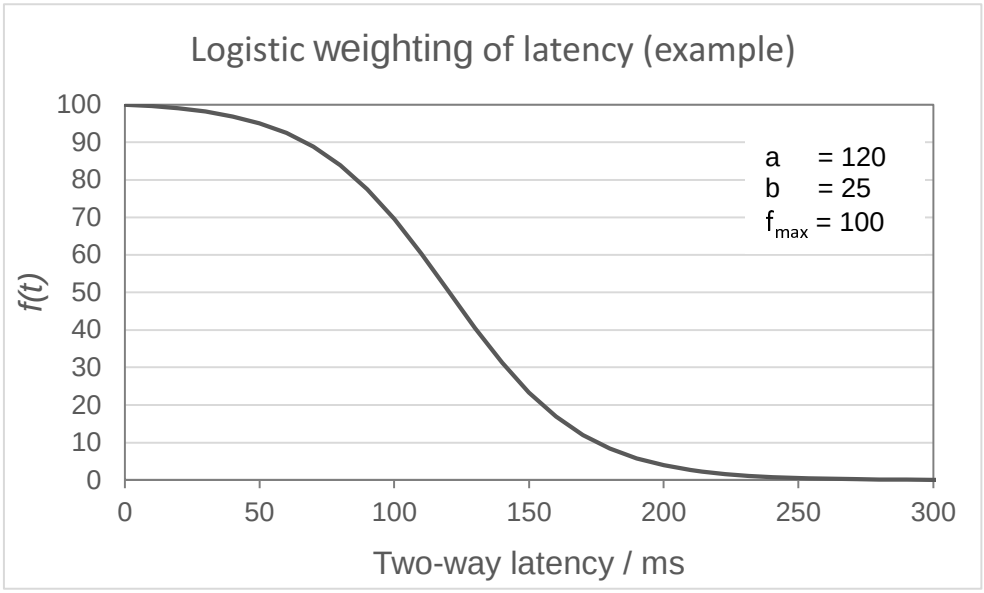

Figure B.1: Latency example showing tail saturation

If seen from an QoE perspective, first, parameter *a* shifts the latency value, where the decrease of perceived interactivity depending on latency starts along the *t*-axis. In case f_0 is close to 1,0, parameter *a* directly defines the latency, where the predicted interactivity is decreased to 50 % of the maximum reachable score. Second, *b* determines the sensitivity of the user, means the range of latency, where the perceived interactivity is decreasing from almost maximum close to zero. Both parameters are depending on the expectation of the user to the application used.

The value $f_P(t, i)$ is computed for each individual packet *i* and its latency *t*. The aggregated interactivity based on latencies I_L for an observation period is the accumulation of all $f_P(t, i)$:

$$I_L = \frac{1}{N}\sum_{i=1}^{N} \frac{f_{max}}{f_0}\left(1 - \frac{1}{1+ e^{-\frac{1}{b}(t_i - a)}}\right)$$

In cases where the per-packet two-way latency cannot be obtained in the test setup or higher order statistics appear more applicable, the principle of the logistic weighting can be applied to the median of the two-way latencies as a single value input into the function.

$$I_{L-CG} = \frac{f_{max}}{f_0}\left(1 - \frac{1}{1+ e^{-\frac{1}{b}(RTT_{MEDIAN} - a)}}\right)$$

$$\text{with } \mathrm{RTT_{MEDIAN}} = \text{Median delay of all packets sent by the application}$$

If using the alternative method, the parameters *a* and *b* may differ from the per-packet application of the sigmoid function.

In addition to latency, it is anticipated that delay variation and amount of disqualified packets contribute to the perceived interactivity too.

To reflect the packet delay variation within the applied delay budget, the formula of PDV as in clause 5.1.3.3.3 is applied to packet delays by considering only the qualified packets according to clause 5.1.3.3.4 and defined as PDV_Q. In the further modelling, the standard deviation across all PDV_Q values is defined as PDV_{sQ}. PDQ is defined in clause 5.1.3.3.4.

Simplified, both indicators are considered as degrading factors D_{PDV} and D_{DQ} by multiplication:

$$IntAct = I_L \times D_{PDV} \times D_{DQ}, \text{ with } \mathrm{D_{PDV}} = 1 - PDV_{SQ} / \mathrm{u}, \text{ and } \mathrm{D_{DQ}} = 1 - \mathrm{v}\ \mathrm{P_{DQ}}$$

The multiplication with the contributors $0 < D_{PDV} < 1$ and/or $0 < D_{DQ} < 1$ will also decrease the maximum value $\mathrm{IntAct_{max}} < \mathrm{I_{Lmax}}$. It means even if latency is very short, in case packets are disqualified the maximum score for perceived interactivity cannot be reached anymore.

EXAMPLE: A median PDV of 50 ms will lead to a degrading factor $D_{PDV} \sim 0{,}9$ in case $u = 55$, and a ratio of disqualified packets of 5 % ($P_{DQ} = 0{,}05$) to a degrading factor D_{DQ} = 0,65 if $v = 7$.

B.2.2 Example high-interactive 'e-Gaming real-time'

The example traffic pattern for emulating high-interactive e-gaming is derived from heavy multiplayer games. This example is also described in Appendix I to Recommendation ITU-T G.1051 [i.23]. It covers an initializing phase (low bitrate) without interaction, a sustainable phase with motion and interaction, the loading of a new game instance as a bitrate peak, a longer sustainable phase of high interaction with up to several hundreds of players and a (low bitrate) trailing phase with fewer players and medium interaction. The set bitrates for the phases where taken from real, demanding multiplayer games and represent also in its relative duration and sequence a real gaming session but compressed into a duration of 10 s.

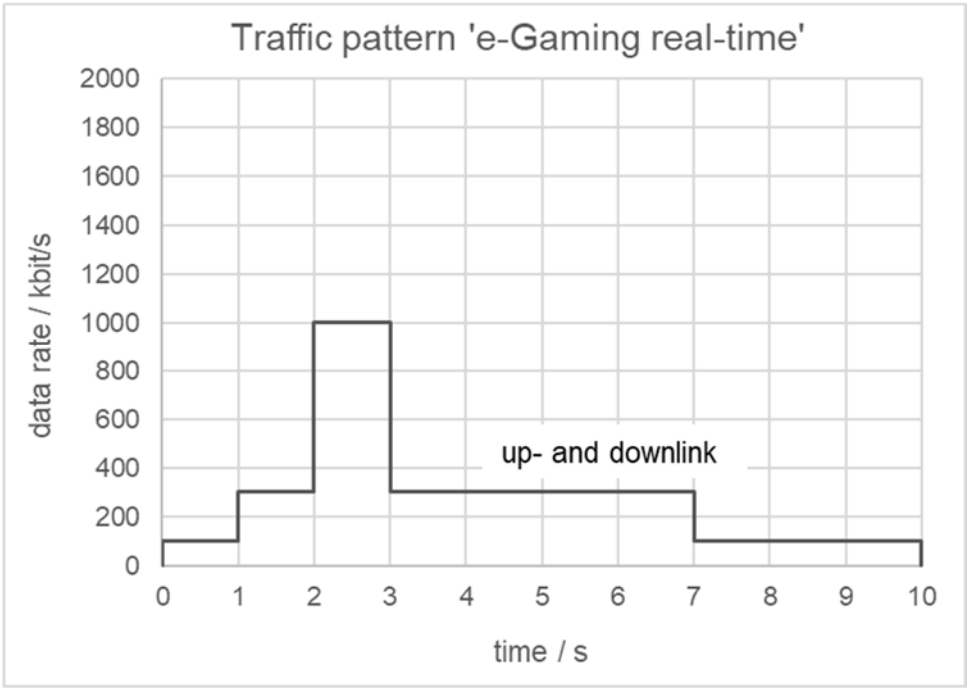

Figure B.2

The chosen packet size is 100 bytes sent in a frequency of 125 to 1 250 packets per second and the pattern is same in uplink and downlink, means each packet is reflected in same size. In ETSI TS 123 501 [i.22], a maximal one-way latency for online real time gaming of 50 ms is defined (5QI class 3), the two-way latency therefore should not exceed 100 ms and forms the delay budget for this test case.

For parametrization of the interactivity model as in clause B.2.1 of the present document, the following principles are used:

- A fluent video stream produces 60 frames per second (fps), a movie 24.

It is assumed that the degradation starts if the channel adds a two-way delay of ~30 ms (two frames delay for 60 fps). Furthermore, a degradation to 60 of 100 is assumed in case of a two-way delay added by the channel of ~60 ms (four frames for 60 fps). These thresholds can be seen as challenging but refer to highly interactive gaming applications in high speed networks as in 5G URLLC.

In addition, it is anticipated that a PDV_{sQ} of 10 ms reduces the interactivity as defined by latency by another 8 % ($D_{PDV} = 0{,}92$) and a ratio of disqualified packets of 5 % reduces the perceived the interactivity by 20 % ($D_{DQ} = 0{,}8$).

The parametrization of the model:

$$IntAct = I_L \times D_{PDV} \times D_{DQ}$$

with $D_{PDV} = 1 - PDV_{SQ}/u$, $D_{DQ} = 1 - v\,P_{DQ}$ and $I_L = \frac{1}{N}\sum_{i=1}^{N}\frac{f_{max}}{f_0}\left(1 - \frac{1}{1+e^{-\frac{1}{b}(t_i-a)}}\right)$

is resulting in:

parameter	f_{max}	*a*	*b*	*u*	*v*
value	100	61	14	300	4

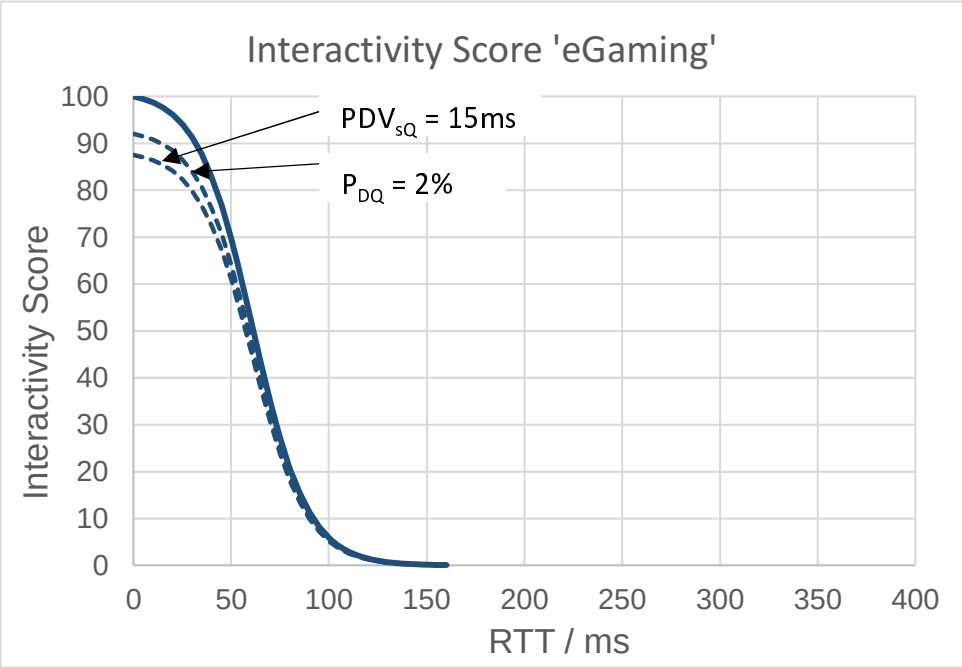

Figure B.3

B.2.3 Void

History

Document history		
V1.1.1	November 2020	Publication
V1.2.1	August 2024	Publication