ITUPublications
Recommendations

International Telecommunication Union
Standardization Sector

Recommendation

ITU-T G.1051 (03/2023)

SERIES G: Transmission systems and media, digital systems and networks

Multimedia Quality of Service and performance – Generic and user-related aspects

Latency measurement and interactivity scoring under real application data traffic patterns

ITU-T G-SERIES RECOMMENDATIONS

TRANSMISSION SYSTEMS AND MEDIA, DIGITAL SYSTEMS AND NETWORKS

INTERNATIONAL TELEPHONE CONNECTIONS AND CIRCUITS	G.100–G.199
GENERAL CHARACTERISTICS COMMON TO ALL ANALOGUE CARRIER-TRANSMISSION SYSTEMS	G.200–G.299
INDIVIDUAL CHARACTERISTICS OF INTERNATIONAL CARRIER TELEPHONE SYSTEMS ON METALLIC LINES	G.300–G.399
GENERAL CHARACTERISTICS OF INTERNATIONAL CARRIER TELEPHONE SYSTEMS ON RADIO-RELAY OR SATELLITE LINKS AND INTERCONNECTION WITH METALLIC LINES	G.400–G.449
COORDINATION OF RADIOTELEPHONY AND LINE TELEPHONY	G.450–G.499
TRANSMISSION MEDIA AND OPTICAL SYSTEMS CHARACTERISTICS	G.600–G.699
DIGITAL TERMINAL EQUIPMENTS	G.700–G.799
DIGITAL NETWORKS	G.800–G.899
DIGITAL SECTIONS AND DIGITAL LINE SYSTEM	G.900–G.999
MULTIMEDIA QUALITY OF SERVICE AND PERFORMANCE – GENERIC AND USER-RELATED ASPECTS	**G.1000–G.1999**
TRANSMISSION MEDIA CHARACTERISTICS	G.6000–G.6999
DATA OVER TRANSPORT – GENERIC ASPECTS	G.7000–G.7999
PACKET OVER TRANSPORT ASPECTS	G.8000–G.8999
ACCESS NETWORKS	G.9000–G.9999

For further details, please refer to the list of ITU-T Recommendations.

Recommendation ITU-T G.1051

Latency measurement and interactivity scoring under real application data traffic patterns

Summary

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

Latency and resulting reactivity must be measured in a scenario that emulates the application and use case to be evaluated. This requires first a data transfer profile (traffic pattern) that is considered as equivalent to the application so that the relevant latency and reactivity can be measured. Second, the resulting influence of latency to a certain application can be described by an interactivity scoring model. This model is not a general one, rather, it is individually scaled for each of the use cases like e.g., e-Gaming or real-time drone control and is focused on scoring transport with a simplified, parametrizable model approach, it does not target individual application behaviours.

History

Edition	Recommendation	Approval	Study Group	Unique ID*
1.0	ITU-T G.1051	2023-03-01	12	11.1002/1000/15468

Keywords

Data traffic patterns, interactivity, latency, round-trip time.

* To access the Recommendation, type the URL http://handle.itu.int/ in the address field of your web browser, followed by the Recommendation's unique ID. For example, http://handle.itu.int/11.1002/1000/11830-en.

FOREWORD

The International Telecommunication Union (ITU) is the United Nations specialized agency in the field of telecommunications, information and communication technologies (ICTs). The ITU Telecommunication Standardization Sector (ITU-T) is a permanent organ of ITU. ITU-T is responsible for studying technical, operating and tariff questions and issuing Recommendations on them with a view to standardizing telecommunications on a worldwide basis.

The World Telecommunication Standardization Assembly (WTSA), which meets every four years, establishes the topics for study by the ITU-T study groups which, in turn, produce Recommendations on these topics.

The approval of ITU-T Recommendations is covered by the procedure laid down in WTSA Resolution 1.

In some areas of information technology which fall within ITU-T's purview, the necessary standards are prepared on a collaborative basis with ISO and IEC.

NOTE

In this Recommendation, the expression "Administration" is used for conciseness to indicate both a telecommunication administration and a recognized operating agency.

Compliance with this Recommendation is voluntary. However, the Recommendation may contain certain mandatory provisions (to ensure, e.g., interoperability or applicability) and compliance with the Recommendation is achieved when all of these mandatory provisions are met. The words "shall" or some other obligatory language such as "must" and the negative equivalents are used to express requirements. The use of such words does not suggest that compliance with the Recommendation is required of any party.

INTELLECTUAL PROPERTY RIGHTS

ITU draws attention to the possibility that the practice or implementation of this Recommendation may involve the use of a claimed Intellectual Property Right. ITU takes no position concerning the evidence, validity or applicability of claimed Intellectual Property Rights, whether asserted by ITU members or others outside of the Recommendation development process.

As of the date of approval of this Recommendation, ITU had received notice of intellectual property, protected by patents/software copyrights, which may be required to implement this Recommendation. However, implementers are cautioned that this may not represent the latest information and are therefore strongly urged to consult the appropriate ITU-T databases available via the ITU-T website at http://www.itu.int/ITU-T/ipr/.

© ITU 2023

All rights reserved. No part of this publication may be reproduced, by any means whatsoever, without the prior written permission of ITU.

Table of Contents

Recommendation ITU-T G.1051

Latency measurement and interactivity scoring under real application data traffic patterns

1 Scope

The scope of this Recommendation is to specify a method to measure continuously data two-way latency and loss for a defined observation period. The basis for this method can be IETF's Two-Way Active Measurement Protocol (TWAMP). The typical symmetrical fixed-rate stream method will be modified to reflect situations in today's telecommunication networks including mobile scenarios and typical interactive use cases. The measurement approach is designed to also cover 5G URLLC configurations.

In this approach, a scalable UDP packet stream is sent 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.

Based on the results of the latency, latency variation and loss measurements, a generic approach for a model describing interactivity as a single figure of merit is developed. Because the measurement results and the interactivity model depend on the chosen traffic pattern emulating an application, there will be clear rules to derive traffic patterns and corresponding model configurations as well as defined traffic patterns and formulas for popular applications (by applying the defined rules). This allows an immediate application of the measurement approach and enables later extension for new, arising applications.

2 References

The following ITU-T Recommendations and other references contain provisions which, through reference in this text, constitute provisions of this Recommendation. At the time of publication, the editions indicated were valid. All Recommendations and other references are subject to revision; users of this Recommendation are therefore encouraged to investigate the possibility of applying the most recent edition of the Recommendations and other references listed below. A list of the currently valid ITU-T Recommendations is regularly published. The reference to a document within this Recommendation does not give it, as a stand-alone document, the status of a Recommendation.

[ITU-T Y.1540] Recommendation ITU-T Y.1540 (2019), *Internet protocol data communication service – IP packet transfer and availability performance parameters*.

[3GPP TS 23.501] 3GPP TS 23.501 (2020), *System architecture for the 5G system (5GS)*.

[IETF RFC 5357] IETF RFC 5357 (2008), *A two-way active measurement protocol (TWAMP)*.

[IETF RFC 5481] IETF RFC 5481 (2009), *Packet delay variation applicability statement*.

[IETF RFC 6038] IETF RFC 6038 (2010), *Two-way active measurement protocol (TWAMP) reflect octets and symmetrical size features*.

3 Definitions

None.

4 Abbreviations and acronyms

This Recommendation uses the following abbreviations and acronyms:

5QI	5G Quality of Service Identifier
ACR	Absolute Category Rating
DL	Downlink direction in mobile connection
HD	High Definition (video)
IP	Internet Protocol
IPDV	Inter-Packet Delay Variation
MOS	Mean Opinion Score
MTU	Maximum Transmission Unit
PDV	Packet Delay Variation
QoS	Quality of Service
RTT	Round-Trip Time
TWAMP	Two-Way Active Measurement Protocol
UDP	User Datagram Protocol
UDPST	User Datagram Protocol Speed Test
UL	Uplink direction in mobile connection
URLLC	Ultra-Reliable Low-Latency Communication

5 Conventions

None.

6 Introduction

This Recommendation describes the technical realization of two-way delay measurements, how to apply this approach to emulate real application traffic and how to obtain the metrics and results from this measurement.

The introduced measurement method considers specific characteristics of mobile networks and, in general, dynamically adjusted and load-dependent networks and connections.

This Recommendation also gives advice on how to derive load and traffic patterns from real applications as well as for adjusting the scalable interactivity model to the target use case or application.

One realization of an on-the-top model to predict interactivity of the tested connection is described in Annex A.

7 Test method to obtain two-way latency

7.1 Approach to obtain two-way latency

A typical approach to obtain two-way latency is sending packets to a reflecting unit in the network and measuring the time between sending and receiving the corresponding packets. Time stamps for each packet from the sending and reflecting unit are required to obtain the key performance indicators (KPIs) defined in this Recommendation.

As an established example, the TWAMP methodology and protocol can be considered, as defined in [IETF RFC 5357]; additional notes can be found in [IETF RFC 6038]. It is based on a UDP packet stream of packets of pre-defined size and frequency. The packets are sent to a reflecting server that sends the packet back to the sending client, where 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 TWAMP protocol according to IETF is defined and – depending on the vendor – supported by infrastructure components such as routers and IP-gateways.

In addition to two-way packet latency, the TWAMP methodology also supports the calculation of one-way packet-delay variation (PDV), separated into uplink and downlink directions, as well as the detection of lost packets based on packet sequence numbers.

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.

Other methods to obtain packet round-trip time (RTT) and one-way delay jitter and packet loss in sufficient temporal resolution can also be used such as that based on the UDPST method as described in [b-IETF] and is available under [b-OB-UDPST].

7.2 Extending the approach to asymmetrical traffic

A pure reflection of received packets by a reflecting unit leads to widely identical and symmetrical data streams in direction to and from the reflecting unit. Each sent packet triggers a reflected packet of the same size. Consequently, packet size and frequency and the resulting data rate are the same. This symmetrical traffic is a special case and does not emulate most real applications, where the traffic is asymmetrical, meaning different in each direction.

Additional notes about the method as specified in [IETF RFC 6038] are focused on achieving absolutely symmetrical traffic; the received packets are reflected exactly back to the sender. Consequently, the amount of data requested to transmit is the same in the send and receive directions.

The definition of additional TWAMP features in [IETF RFC 6038] has already foreseen the possibility to define the size of the reflected packets[1].

A different size of the reflected packet generates asymmetrical traffic; the reflected packet stream can result in a higher or lower bitrate. It must be noted that this way the asymmetrical traffic is only achieved by varying the packet size, not the packet frequency.

There can be implementations realized, where also the packet frequency varies in either direction. A received packet at the reflecting unit can trigger sending multiple packets instead of one. It depends on the implementation and the target traffic pattern to be emulated whether these packets are sent in an equidistant way or as bursts.

Vice versa, there might also be implementations where multiple received packets at the reflecting unit trigger less or just one packet to be sent back to the client side. Also, in this case the targeted traffic pattern defines the parameters of asymmetry.

If the reflected packet must carry a higher number of bytes than the sent one, the packet is extended by random bits. This is the more usual case (uplink rate < downlink rate). If the reflected packet must be smaller, the payload is cut to the target size, where the lower limit is the header information of the packet for its identification.

Alternative methods as UDPST as described in [b.1] also offer techniques to realize different load per direction resulting in a desired asymmetrical traffic pattern.

[1] The implementation is not part of [IETF RFC 6038] and is left to the implementor. Preferably, the implementation carries the information of how many octets should be reflected in each packet. Thus, the asymmetry of the traffic can be defined per packet as highest granularity.

7.3 Guidelines to derive traffic patterns from real applications

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

First, the traffic created by several representative real applications should be analysed. The IP trace should be recorded in the different phases of application usage, for example 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.

7.3.1 Temporal structure of traffic patterns

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 s 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 asymmetrical traffic scenarios, the uplink and downlink bitrates can differ and have a different proportion for each segment.

7.3.2 Controlling data rate by packet size and frequency

The packet stream realized e.g., by TWAMP as in [IETF RFC 5357] and [IETF RFC 6038] is considered an emulation of a traffic pattern of a real application. This load is not only defined by an average or short-term bitrate but rather by the underlying packet sizes and frequency. Packet size and frequency can be different for individual applications even the resulting bitrate is the same. Some applications may send small packets in higher frequency, while others send larger packets but in lower frequency.

Therefore, 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 e.g., by TWAMP as in [IETF RFC 6038] (~65 000 bytes). Within this range, the packet size and frequency should be set to resemble the real application traffic.

Packet sizes exceeding one maximum transmission unit (MTU) will be split into multiple MTUs transmitted and received sequentially. The receiving side will then assemble the incoming MTUs to one UDP packet. It should be noted that the latency and the corresponding PDVs are calculated based on the reception of the latest MTU forming this individual packet. Furthermore, a packet is counted as lost if a single MTU is lost or erroneous.

This consideration of oversized packets will result in data streams close to a real application. An oversized packet is split into several MTUs but these MTUs are transferred directly one after another as a bulk of MTUs. After transmitting, there can be a pause until the next oversized packet is split and transmitted. The more bursty data traffic on lower layers is closer to resembling the reality for those applications.[2]

One additional restriction applies to the packet frequency if TWAMP as in [IETF RFC 5357] is used: The defined packet frequency applies for uplink and downlink traffic within one segment, because of the round-trip nature of the measurement. Packets are only reflected at the server side, not multiplied nor discarded. Thus, the packet frequency in both directions is identical.

[2] A typical example application is UHD video streaming, where single frames are packetized but exceed the MTU size and are split. Nevertheless, the entire packet (video frame) is considered as lost if the transport of a single MTU fails.

8 Latency test results and metrics

8.1 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.

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

– Delay *D(i)* 50th percentile (median)

– Delay *D(i)* 10th percentile (approximation for the shortest reachable latencies in practice)

An arithmetic average as 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.

8.2 Packet delay variation

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

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

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.

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 (approx. maximum)

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.

8.3 Inter-packet delay variation

In line with [IETF RFC 5481], the inter-packet delay variation (IPDV)[3] 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.

[3] Please note that IPDV is also used as an abbreviation for IP-packet delay variation in other contexts.

This vector *IPDV*(*i*) is used to calculate the:

– $IPDV_{>0}$ 34.1th percentile (standard deviation under assumption of normal distribution)

– $IPDV_{>0}$ 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:

– $IPDV_{\mathrm{CS}}$ one-way (client to server)

– $IPDV_{\mathrm{SC}}$ one-way (server to client)

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

8.4 Lost packets

Packets which are not received on the IP-interface of client or server side are counted as lost.

This covers more than lost packets that are actually lost, 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. Lost packets can be determined at the reflecting server side as well as at the receiving client. Lost packets can be reported for either one-way separately or two-way.

– Erroneous packets: Packets that were corrupted after arriving back at the client device.[4]

An overall indicator as the ratio of lost packets P_{L} can be computed simply by:

P_{L} = *number of lost packets / number of all packets sent by the application*

The packet loss P_{L} can also be calculated per direction separately if the underlying method allows it.

$P_{\mathrm{L\text{-}CS}}$ = *number of lost packets at server side / number of all packets sent by the client application*

$P_{\underline{\mathrm{L\text{-}SC}}}$ = *number of lost packets at client receiving side / number of all packets sent by the server*

9 Interactivity prediction model and related test case definition

9.1 Principles of the generic interactivity model approach

Obtaining latency, PDV, IPDV and packet loss opens the possibility to create a prediction model for perceived interactivity for individual applications or application classes. An application class could be, for example, HD video chat or online gaming, where the effect of delay, jitter and loss on perception is comparable in between the individual realizations of the application.

The basic concept of scoring interactivity is that the latency and number of lost and unusable packets determine the interactivity perceived by a user. Two-way packet latency gives information on how fast a response to an action originated at the client user device is received back at the device side. In

[4] Please note that 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.

addition, disqualified packets are missing information for the user's application. They can be lost packets as in clause 8.4 or packets considered as discarded by exceeding a delay limit. Those lost and discarded packets are defined as disqualified packets.[5] Whether disqualified packets can be interpolated by the application or only lead to temporary distortions, pausing while using the application or even stopping the application completely depends on the application itself and its implementation.

To receive results for latency and ratio of disqualified packets as close to a real application as possible, the packet stream – especially in data-rate and traffic pattern – should emulate the targeted application in use.

Latency of the received packets and number of disqualified packets determine the perceived interactivity, but the influence on the perceived interactivity of an application or use case depends on the target application and the expectation of the user.

Consequently, there will not be one single prediction model for interactivity, but rather individual ones for different applications. Considering the huge number of potential applications, a generic and scalable approach is developed and described in this Recommendation.

Therefore, the following principles are anticipated:

– Same input parameter types for derived interactivity models

– Same computational structure of the interactivity model

– Classification of applications and use cases in application types or groups having same or similar expectations of the user

– Parametrization of the computational model structure to one specific application type or group

Considering the principles above, a fast realization of prediction models is achievable and because of clear application grouping and a transparent approach, the results stay comparable within and across application groups. One example of a model for the computation of an interactivity score based on the principles described above is given in Annex A.

9.2 Parametrization of an interactivity model

9.2.1 Subjective testing

The most obvious approach to parametrization is the use of data obtained by subjective testing of the application that should be predicted by the interactivity model. Human subjects have to use the application in its interactive stage under different defined channel conditions regarding latency, jitter and loss and to score the perceived interactivity. Based on the results, the best fitting parameters for the above-described model can be derived.

The advantage of this approach is the direct link to human scores of a real application and thus the ability to correctly reflect the priority/weight of different network degradations. Disadvantages are that the application has to exist and that the outcome of the test refers to this dedicated application and it is hardly possible to generalize unless also evaluating how other competing applications behave, which does not have to be done with fully subjective tests. Furthermore, the ambition of a test user in an experimental set-up may not be fully comparable with a real use of the application, where the interests of the user are different.

[5] Considering packets exceeding the delay budget as unusable and discarded reflects real-time applications, where delayed packets cannot be considered for e.g., rendering a media stream.

9.2.2 Best practice parametrization

Best practice parametrization models perceive interactivity based on anticipated thresholds for a given application or use case. These anticipated thresholds are derived from expected performance in certain environments while using the application. Basic considerations can be at which latency a perceived degradation of latency starts (e.g., an interactivity score of 90 of 100) and at which latency the perceived interactivity drops significantly (e.g., an interactivity score of 60 of 100, would relate to MOS ~ 3 in a five-point absolute category rating (ACR) test) and/or where the application becomes unusable in practice. In addition, the influence of packet loss and serious packet delay variation should be estimated. This estimation considers known or assumed packet loss and buffering strategies. Examples for best practice parametrization are shown in Appendices I and II..

9.3 Guidelines for application grouping

In principle, an individual traffic pattern and interactivity model can be defined for each application or use case. However, there are applications and use cases resulting in very similar patterns and model parameters. For comparability and practical use, it is helpful to test applications and use cases under the same conditions if the traffic shape and/or model parameters are similar.

Traffic pattern similarity should be given in

– Temporal structure of traffic pattern;

– Applied bit rates.

Model parameter similarity should be given in

– Delay budget for the application;

– Perceptual influence of packet latency;

– Degradation by packet latency variations and especially disqualified packets.

The use of the same defined traffic pattern and set of model parameters for different applications is not limited to similar appearance or use; applications can be treated the same although they have a different use, as long as they share the similarities above.

9.4 Guidelines for defining test cases

The results obtained by the test method, such as two-way latency, PDV, IPDV and packet loss, depend on the traffic emulated, for example, on bitrate and packet size. Therefore, the measurement of latency must take place under equivalent traffic conditions as in the target application for which latency scores should be obtained.

If the measurement results will be used to estimate the interactivity of an application type, the parametrization of this model must also consider the characteristics of the targeted application such as sensitivity to packet loss, defined delay budget and dependency on latency.

Considering this, a test case for deriving a prediction of interactivity is mainly defined by the used traffic or bitrate profile including packet size and frequency, and – if applied – the model parameters to estimate perceived interactivity. A packet with a two-way latency exceeding the defined delay budget is considered as disqualified, meaning discarded due to late arrival by the emulated application.

Limits for latency of different application classes, grouped by the standardized 5QI value, are given by [3GPP TS 23.501]. These limits apply to network delays. They are preferable to use if the emulated application matches, otherwise best practice assumptions or actual discarding limits of the applications are practicable.

Examples of traffic patterns and corresponding model parameters for a model as described in Annex A are presented in Appendix I.

Annex A

Computational structure of a generic interactivity model approach

(This annex forms an integral part of this Recommendation.)

A.1 Introduction

This annex describes a model for estimation of perceived interactivity. As perceived interactivity strongly depends on the application class, the computational model must be parametrized for dedicated application classes and their demands.

The model is based on the obtained two-way latency per packet from simulated data traffic in a network or transport centric approach.

The basic idea of this model realization is a simple consideration of the application's client and server and its scalable construction across many different application cases. Even though this will not perfectly match each individual application, its main advantage is the comparability between individual settings and modelled test cases due to the same modelling approach being taken.

A.2 Modelling approach for perceived interactivity of interactive applications

The basic assumption of modelling perceived interactivity is its monotonous dependency on data latency. The shorter the data transport time is, the shorter the response time in an interactive application is and the more interactive the use of the application is perceived to be.

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

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

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

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

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

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

where t is the packet latency in milliseconds, i is the indicator of the packet, a defines the shift of $f_\mathrm{P}(t, i)$ directly on the t-axis in milliseconds and b defines the gradient of $f_\mathrm{P}(t, i)$, where larger values of b make $f_\mathrm{P}(t, i)$ less steep.

The scaling factor f_0 guarantees the maximum score value at $t = 0$. In case, the upper saturation area of the $f_\mathrm{P}(t, i)$ starts in the range of $t > 0$, the f_0 will be close to 1.0. If $f_0 \to 1$, the parameter a defines the latency value where the perceived interactivity has fallen to almost 50% of f_max.

In cases where per-packet two-way latency can be obtained, the logistic function can be applied to each individual packet latency. Here, the value $f_\mathrm{P}(t, i)$ 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 average of all $f_\mathrm{P}(t, i)$:

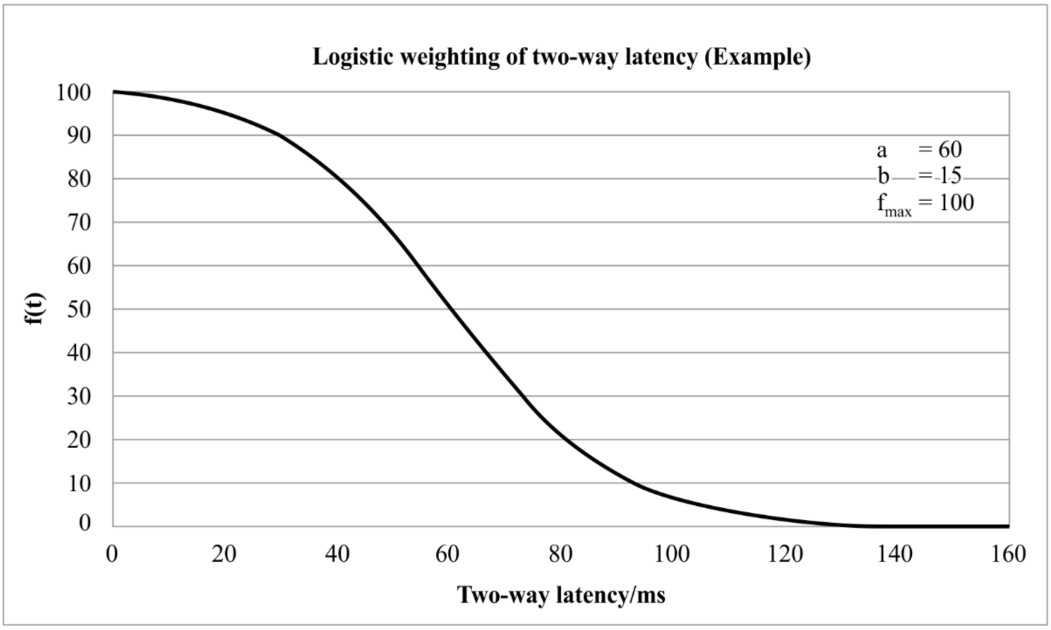
$$I_L = \frac{1}{N}\sum_{i=1}^{N} \frac{f_{max}}{f_0}\left(1 - \frac{1}{1 + e^{-\frac{1}{b}(t_i - a)}}\right)$$

In cases where the per-packet two-way latency cannot be obtained in the test set-up 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.

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

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

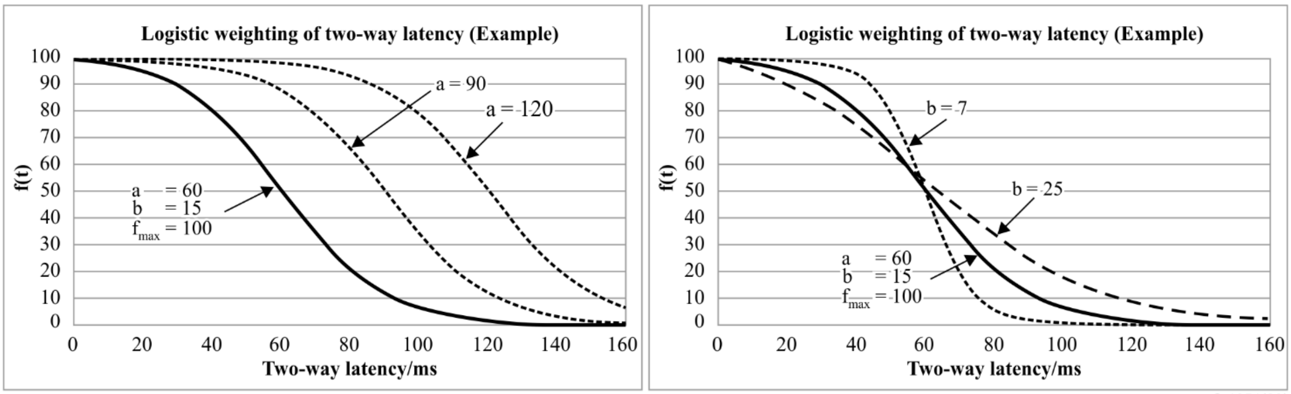
If using the alternative method, the parameters *a* and *b* may differ from the per-packet application of the sigmoid function.



G.1051(23)

Figure A.1 – Example of a logistic weighting function for two-way latency

If seen from an quality of experience (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 to close to zero. Both parameters are depending on the expectation of the user to the application used.



G.1051(23)

Figure A.2 – Parametrization of the logistic weighting function for two-way latency

A.3 Consideration of lost and discarded packets and packet delay variation in the model

A.3.1 Disqualified packets

The defined packet loss P_L reports packets which are not received and cannot be used by the application. In addition to that, a real application will also discard packets which are heavily delayed and are not received in time. This maximum accepted delay for packets depends on the application and e.g., internal queuing and buffering mechanisms. In general, the more real-time capability an application has to provide, the shorter the acceptable delay. Packets received later than the maximum delay are discarded by the application and would not be usable for a real and running application, for example, for media rendering.

For predicting a perceived interactivity, discarded packets count as lost packets (P_L) and are counted cumulatively as disqualified packets. This covers 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. Lost packets can be determined at the reflecting server side as well as at the receiving client. Lost packets can be reported for either one-way separately as well as for two-way.

– Erroneous packets: Packets that were corrupted after arriving back at the client device.

as considered in clause 8.4 as lost packets, and additionally

– Discarded packets: Packets that were received back after a pre-defined timeout at the client device. This timeout, also called delay budget, emulates discarding by a real application because of too long a delay. The timeout is specified and defined according to the maximum acceptable latency for the target application and forms one parameter of emulation of the application.

From an application's point of view, no differentiation is needed between the individual causes of not considering a packet as received. An overall indicator such as the ratio of disqualified packets P_{DQ} is seen as sufficient at application level.

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

As a consequence, all packets considered as usable by the application are defined as qualified packets.

A.3.2 PDV considering delay budget

In clause 8.2, the packet delay variation (PDV) is defined. The PDV is derived from the received packet stream without considering discarding due to the applied delay budget. As for prediction of the perceived interactivity a delay budget is applied, the PDV of the incoming packet stream does not reflect anymore the packet delay variation to be considered by the emulated application.

To reflect the packet delay variation within the applied delay budget, the formula of PDV is applied to packet delays considering only the qualified packets according to clause A.3.2.1 and defined as PDV_{Q}. In the further modelling, the standard deviation across all PDV_{Q} values is defined as PDV_{sQ}.

The PDV_{sQ} describing delay variation from client to server and return can be calculated based on two-way latencies, where the time difference between sending a packet and receiving back its reflection is measured.

If the two-way PDV cannot be obtained, alternatively the standard deviation of PDV or IPDV can be computed individually per link direction.[6] Here, IPDV and resulting $IPDV_{\mathrm{sQ}}$ are measured for either direction separately and the average of the standard deviations of IPDV in the uplink and downlink direction is considered as substitute of two-way PDV_{sQ}. If using the per-link computation, the parameter u may differ from that defined for two-way PDV_{sQ}.

[6] This separation into IPDV per link is also required in case of asymmetrical traffic patterns using different packet frequencies in each direction. There is no one-to-one relation between sent and received packets which is the pre-condition for a two-way PDV or two-way IPDV.

A.3.3 Consideration of disqualified packets and two-way PDV in the interactivity model

In addition to latency, it is anticipated that delay variation and number of disqualified packets also contribute to perceived interactivity. To simplify, both indicators are considered as degrading factors D_{PDV} and D_{DQ} by multiplication:

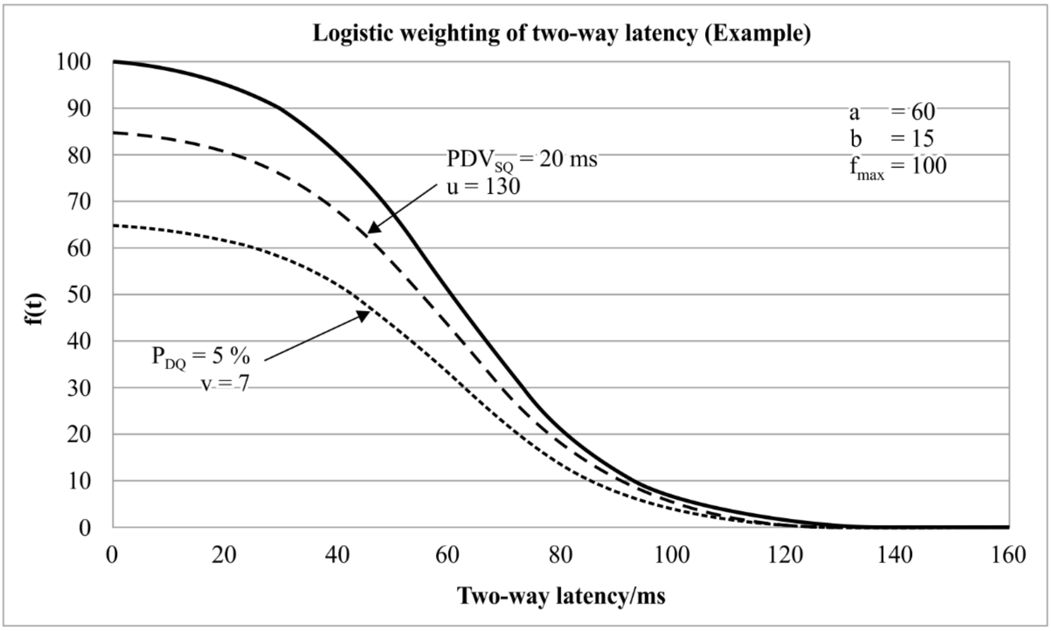
$$IntAct = I_{\mathrm{L}} \times D_{\mathrm{PDV}} \times D_{\mathrm{DQ}}, \text{ with } D_{\mathrm{PDV}} = 1 - PDV_{\mathrm{sQ}} / \mathrm{u}, \text{ and } D_{\mathrm{DQ}} = 1 - v\, P_{\mathrm{DQ}}$$

In the expression above, PDV_{sQ} is the standard deviation across all $\mathrm{PDV_Q}$ values, P_{DQ} is the ratio of disqualified packets, and u and v are parameters that determine the impact of the degradations on interactivity. Alternatively, PDV_{sQ} can also be used as descriptor for the standard deviation of $IPDV_{\mathrm{Q}}$.

The multiplication with the contributors $0 < D_{\mathrm{PDV}} < 1$ and/or $0 < D_{DQ} < 1$ will also decrease the maximum value $IntAct_{\mathrm{max}} < I_{\mathrm{Lmax}}$. It means even if latency is very short, if packets are disqualified the maximum score for perceived interactivity can no longer be reached.

For example, a median PDV_{sQ} of 20 ms will lead to a degrading factor D_{PDV} ~ 0.85 if u = 130, and a ratio of disqualified packets of 5% (P_{DQ} = 0.05) to a degrading factor D_{DQ} = 0.65 if v = 7.

The influence of PDV_{sQ} and P_{DQ} to the weighting function is illustrated in Figure A.3.



G.1051(23)

Figure A.3 – Example of influence of PDV_{sQ} and P_{DQ} to the logistic weighting function

Generally, the above-mentioned formula is applied to the entire observation period (e.g., 10 s). To result in a higher temporal granularity, the formula can also be applied to shorter sub-intervals of 1 s, for example.

A.4 Conclusion

The described computational formula is a simplified, generic approach to retrieve perceived interactivity from latency measurements with the advantage of transparent scalability. The logistic weighting function to estimate a perceived interactivity value from two-way latency measurements ensures a monotonous behaviour and can model perceptual saturation areas at both boundaries of the scale similar to typical perceptual weightings. Degrading effects of delay jitter (derived from PDV_{sQ}) and disqualified packet loss (P_{DQ}) are considered by simple multiplicative scaling factors. Multiplicative scaling retains the monotonous behaviour and does not change the saturation areas and their cut-off positions.

There may be more accurate formulas for perceived interactivity and individual model structures for dedicated applications, but those definitions require more investigations and are for further study.

Appendix I

Example generic traffic patterns and model parameters according to Annex A

(This appendix does not form an integral part of this Recommendation.)

I.1 Principle of application emulation and model parameters

There are many individual traffic patterns when using an application. They depend on the application itself, the phase in the application (e.g., the gaming scene) and the interaction by the user.

Instead of exactly simulating an individual pattern, a more generalized approach is chosen. At first, applications are classified under consideration of similar use and resulting load, e.g., video chat applications or gaming. For those classes of applications, an analysis of real traffic when using the most common applications of each class was made. The derived patterns do not just represent an average bitrate, rather they emulate some statistical characteristics of the real patterns such as the relative occurrence of data rates (e.g., 50% ~100 kbit/s, 30% ~500 kbit/s and 20% ~2 Mbit/s) as well as typical temporal behaviour in a simple way by e.g., incorporating peaks or longer steady transport.

This principle will load the network statistically in a similar way to a real application of this class and leads to latencies as they can be expected when running this type of application.

These example traffic patterns and the resulting two-way latency values, the measured PDV and packet loss can also be used to estimate a perceived interactivity for this application class by applying a computational interactivity model as described in Annex A.

The model as in Annex A is based on measurements on the transport layer; the measurements do not consider the application itself with respect to specific treatments of transport problems such as packet loss concealment or predictive media rendering, for example, as in video or gaming. The examples given in this appendix are parametrized by applying limits and thresholds defined by 3GPP for individual use cases and might be more challenging than subjective experience including real applications with error treatment.

I.2 Examples for application emulation and interactivity score computation

I.2.1 Example high-interactive 'e-Gaming real-time'

The example traffic pattern for emulating high-interactive e-gaming is derived from heavy multiplayer games. It covers an initializing phase (low-bit-rate) without interaction, a sustainable phase with motion and interaction, the loading of a new game instance as a bit rate peak, a longer sustainable phase of high interaction with up to several hundreds of players and a (low-bit-rate) trailing phase with fewer players and medium interaction. The set bit rates for the phases were 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.

The chosen packet size is 100 bytes sent in a frequency of 125 to 1250 packets per second and the pattern is the same in uplink and downlink, meaning each packet is reflected in the same size. In [3GPP TS 23.501] 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.

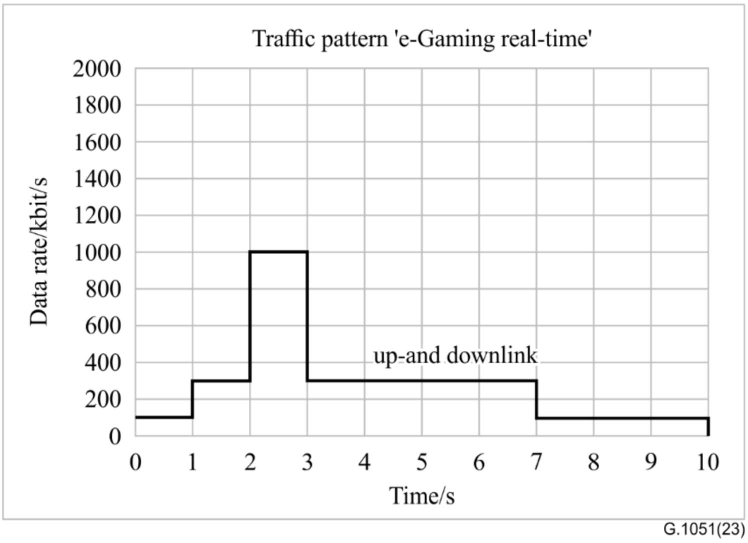


Figure I.1 – Example of traffic pattern 'e-Gaming real-time'

For parametrization of the interactivity model as described in clause 9.2.2, the following principles are used: A fluent video stream produces 60 frames per second (fps), a movie 24. It is assumed that 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 for the interactivity score 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 30 ms reduces the interactivity as defined by latency by another 10% ($D_{\mathrm{PDV}} = 0.9$) and a ratio of disqualified packets of 5% reduces the perceived interactivity by 20% ($D_{\mathrm{DQ}} = 0.8$).

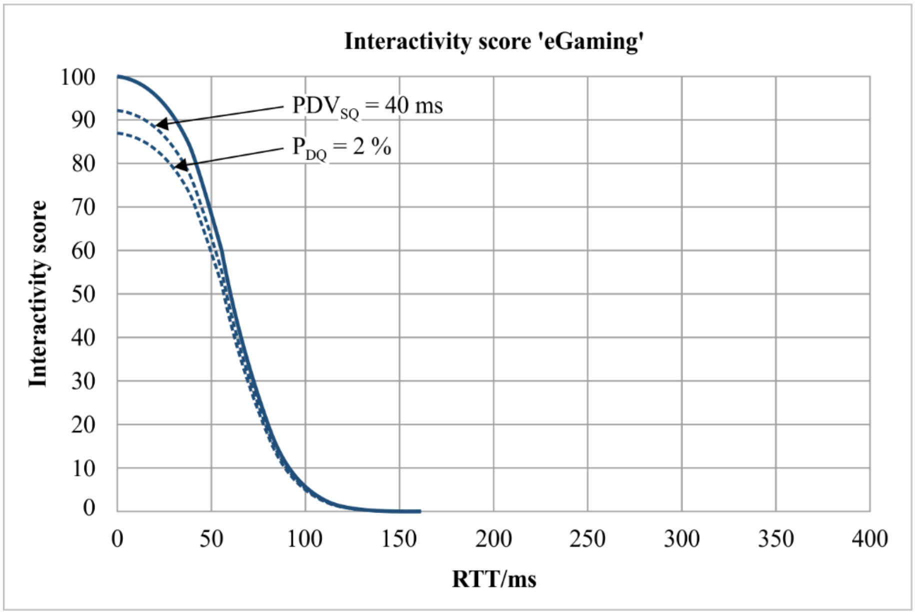
The parametrization of the model:

$$IntAct = I_{\mathrm{L}} \times D_{\mathrm{PDV}} \times D_{\mathrm{DQ}}$$

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

is resulting in:

Parameter	$f_{\max}$	*a*	*b*	*u*	*v*
Value	100	61	14	120	4



G.1051(23)

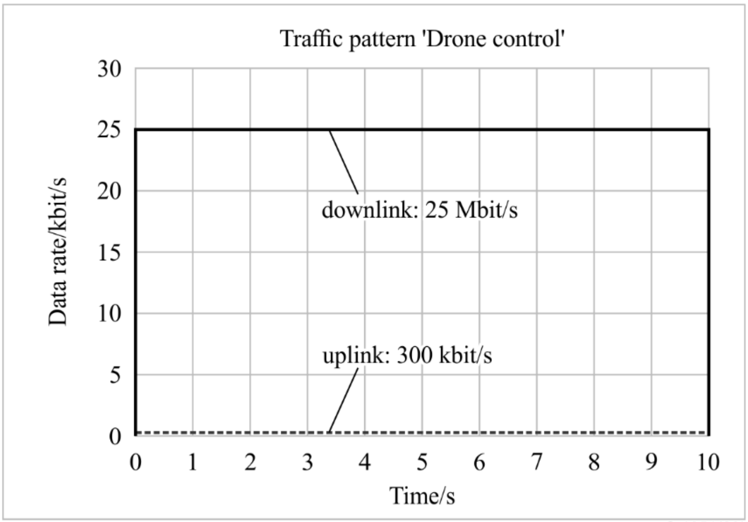
Figure I.2 – Example parametrization for real-time gaming according to Annex A

I.2.2 Example 'Remote drone control'

An example traffic pattern for emulating remote drone control through video illustrates steady but highly asymmetrical traffic, where a low-bit-rate stream of control commands is transferred in uplink to the drone, while the controlling device receives a continuous high bitrate video stream. There are no individual phases with different bit rates; the profile is seen as constant over the observation period. For simplification, the frequency of small uplink control packets and received images is set to the same value. This enables the use of the packet reflection approach as described in [IETF RFC 6038].

The bit rate profile for real-time drone control consists of a 10 s constant 300 kbit/s in uplink and 25 Mbit/s in downlink as defined in use-case scenario 4 ('Remote unmanned aerial vehicle controller through 4k HD video') in [b-3GPP TS 22.125].

Although for the example 'Drone control' the downlink video resolution is chosen to be 4K HD, lower video rates are possible. The example emphasizes the large difference in scale that is possible between uplink and downlink data rate, e.g., 20 times more traffic.



G.1051(23)

Figure I.3 – Example of traffic pattern 'drone control'

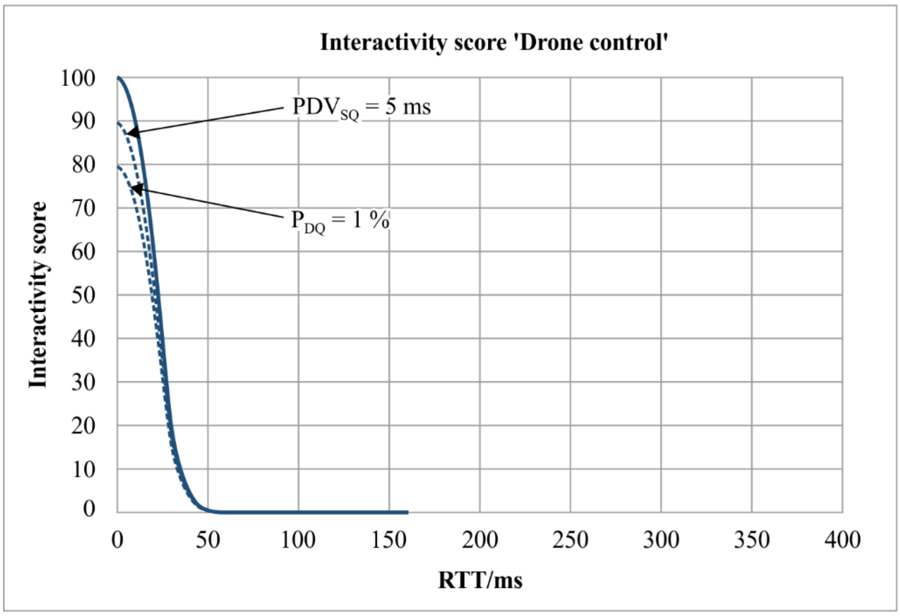
The chosen packet size in uplink is 300 bytes sent in a frequency of 125 packets per second and the corresponding packet size in downlink is 25 000 bytes, meaning each packet is reflected in a much larger size. In [b-3GPP TS 22.125] a maximal one-way latency for remote control of 20 ms unmanned aerial vehicle to a terminating unmanned aerial vehicle is defined. The two-way latency therefore should not exceed 40 ms and this value forms the delay budget for this test case.

For parametrization of the interactivity model as described in clause 9.2.2, the following principles are used: A commercial consumer drone reaches 60 km/h average maximum velocity. This results in a flying distance of ~1.6 cm/ms. A beginning degradation (90 of 100 for the interactivity score) can be anticipated at a delay of visual feedback of 10 ms (flying distance ~16 cm), a severe degradation (60 of 100 for the interactivity score) can be assumed for 20 ms and the begin of an unusable two-way delay at > 30 ms (flying distance > 50 cm).

The additional influence of PDV_{sQ} and disqualified packets is also more severe than for e-gaming. It is anticipated that a PDV_{sQ} of 5 ms reduces the interactivity as defined by latency by another 10% (D_{PDV} = 0.9) and a ratio of disqualified packets of 1% reduces the perceived interactivity already by 20% (D_{DQ} = 0.8).

The parametrization of the model is resulting in:

Parameter	f_{max}	*a*	*B*	*u*	*v*
Value	100	22	6	40	20



G.1051(23)

Figure I.4 – Example parametrization for drone control according to Annex A

I.2.3 Examples 'Interactive remote meeting' and 'Video chat HD'

An example traffic pattern for emulating interactive remote meetings with sharing visual information such as live camera streams, desktop sharing and presentations (with change of content) is emulated by steady, sustainable phases of 500 kbit/s symmetrical traffic and two short intervals of 2000 m peaks, where content change (e.g., new slide) happens. This can be seen as a simplified traffic shape typical for several popular remote meeting tools. This example applies asymmetrical traffic, where one peak is applied in uplink, the other one in downlink. It emulates short peaks in either direction.

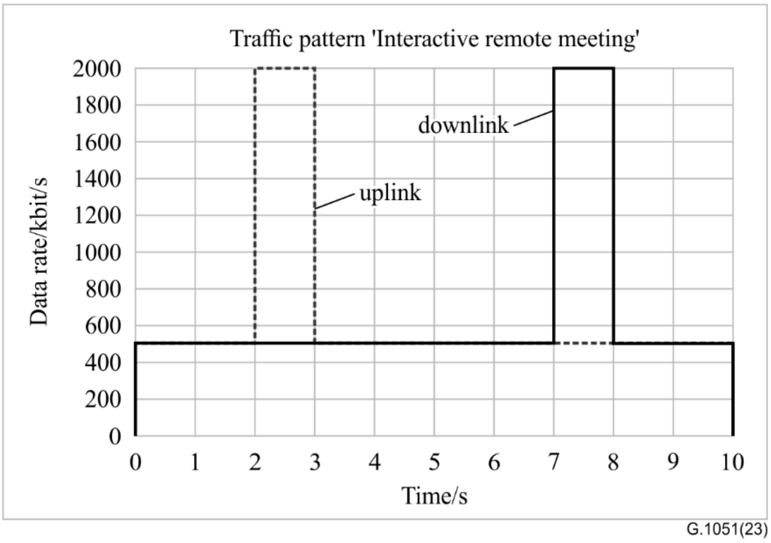


Figure I.5 – Example of traffic pattern 'interactive remote meeting'

The chosen packet size is 1000 bytes, which is typical for video content. The packet sending frequency is then derived to be 62 packets/s during the sustainable phases and 250 packets/s in the high bit rate phases. An example for a 'video chat HD' application is a similar use case, but it relies on live camera video and transmitting the participant videos in high resolution and the main focus is visual interaction and feedback between people. According to real video chat applications, there is a short initial (set-up) phase followed by a sustainable phase where the video link is established.

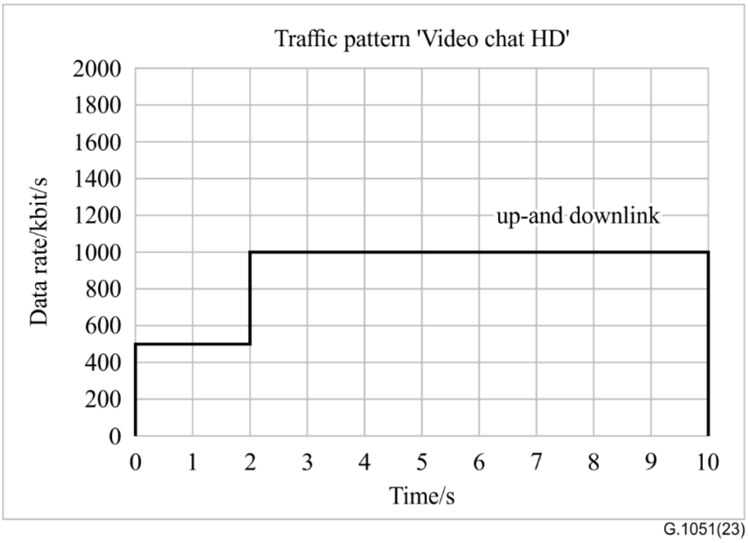


Figure I.6 – Example of traffic pattern 'video chat HD'

The chosen packet size is again 1000 bytes, which is typical for video content. The packet sending frequency is then derived to be 62 packets/s during the sustainable phases and 125 packets/s in the high bit rate phases. The traffic pattern is symmetrical, meaning the packet size and frequency are the same for uplink and downlink.

The applicable delay budget is identical for both types of video application. In [3GPP TS 23.501], a maximal one-way latency for conversational video of 150 ms is defined (5QI class 2), the two-way latency therefore should not exceed 300 ms and this value forms the delay budget for this test case.

These two examples of applications and traffic shapes have similar constraints regrading perceived interactivity and as an example, the applicable interactivity uses the same parameters for both applications and traffic shapes.

For parametrization of the interactivity model as described in clause 9.2.2, the following principles are used: While running a video chat with interactions and feedback, a two-way delay of 100 ms can be seen as the length of a spoken syllable, a beginning degradation is anticipated for this range (90 of 100 for the interactivity score). The interactions become very difficult in case of a response delay > 250 ms.

It is further expected that while using an interactive remote meeting, a user may have more relaxed expectations on packet delay variations than for e-gaming but is more sensitive in case of lost or disqualified packets because video meetings are usually based on unreliable transmission and missed frames are visible as image distortions.

To reflect this, it is anticipated that a PDV_{SQ} of 25 ms reduces the interactivity as defined by latency by another 10% ($D_{\mathrm{PDV}} = 0.9$) and a ratio of disqualified packets of 1% reduces the perceived interactivity already by 30% ($D_{\mathrm{DQ}} = 0.7$).

The parametrization of the model is resulting in:

Parameter	f_{max}	*a*	*b*	*u*	*v*
Value	100	215	50	150	30

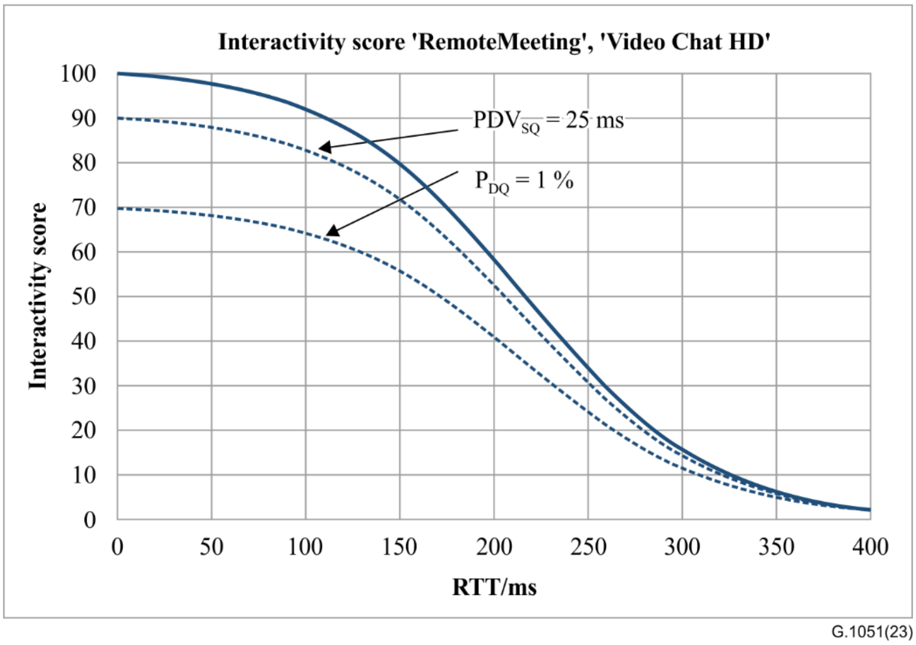


Figure I.7 – Example parametrization for remote meeting and video chat according to Annex A

Appendix II

Parameter implementation according to Annex A based on subjective results for cloud gaming considering a real example application on client and server

(This appendix does not form an integral part of this Recommendation.)

II.1 Introduction

Subjective tests show that subjects are less sensitive to delay and delay jitter in a real gaming situation than the 3GPP thresholds define. There can be many reasons, but in subjective testing the actual application on server and client are considered.

Even though the model described in Annex A is based on measurements on the transport layer and the measurements do not consider the application itself, the model in Annex A can be parametrized to approximate the subjective perception as in subjective tests.

The given examples in this Appendix II are parametrized by a real gaming situation, while the parameters in Appendix I are based on QoS limits. The subjects played a CS_GO FPS game of 90 seconds match rounds over a Steam Link connection. Consequently, the example parameters in this appendix differ from those given in Appendix I, because the model approximate subjective scores include a real application and its gaming situation.

II.2 Traffic patterns

II.2.1 Packet pattern temporal structure

The recorded packets show that the pattern is mostly tied to the video and audio frame rate. The 120 Hz frame rate and the 50 Hz audio frame rate have the shortest common period of 100 ms, when 12 videoframes and five audio frames are transmitted. This has been noticed from the fact that periodical video and audio frame bursts are transmitted one after each other with a common denominator period of 100 ms. However, the common period can be approximated quite accurately with a 40 ms period comprising five video frames and two audio frames.

The packets recorded on DL show that a video frame is represented by six packets and audio frame by two packets when now network impairments was applied. The last packets of each of these two packet burst types (video frames, audio frames) vary typically between 500 and 800 bytes, and the other packets are of full MTU size. The total size should depend on codecs, resolution and compression rate.

From both DL and UL recorded packets it can also be seen that there is in both directions an acknowledgement mechanism causing transmission of smaller single packets in between bursts.

In addition to the transmission of the apparent acknowledgements, the recorded UL packets show that the frame rate is present, and it shows that user input is transmitted at this rate.

UL packet recording also shows a packet appearing approximately at 50 Hz, which likely is audio input from the user's microphone.

It should be noted that modern algorithms for prioritizing traffic in network nodes analyse the pattern of the traffic. Therefore, in order to closely mimic the real cloud game service, it is important that the traffic characteristics reflect the presence of audio and video streams and the extra packets indicating a real-time transmission.

Table II.1 presents the DL and UL packets patterns basic characteristics represented by packet size and number, based on the analysis of the recorded UL and DL packets described above.

Table II.1 – Packets patterns characteristics

Network condition: No degradations				**Network condition: worst case**			
UL		**DL**		**UL**		**DL**	
Packets/s	kbit/s	Packets/s	kbit/s	Packets/s	kbit/s	Packets/s	kbit/s
~150	~140	~1100	~10000	~700	~620	~600	~3500

As already mentioned above, dividing the packet pattern into temporal sub-segments of 40 ms is expected to be accurate enough. However, to handle the different number of packets per link direction the protocol of choice needs to support, or be extended to support, the concept of bursts (in this case frames) instead (or in addition to) of the single packet concept. This results in two-way measurements such as RTT on the frame burst level instead of the single packet level, but only for the first packet on the respective link.

II.3 Guidance for adaptation

II.3.1 Number of adaptations per link

The different packet patterns for DL can be created by adjusting the number of packets representing the video where the traffic is represented by one to five packets of MTU size. Consequently, the DL can be represented by five patterns.

Packet loss (PL) affects the packet patterns by increasing packets per second and bit/s from the server or from the client. Only a high PL significantly affects the amount of extra traffic, but the interactivity reaches very low subjective interactivity at far lower packet losses and thus the PL impact on the packet pattern can be ignored. The exception would be when RTT is low enough (depending on service's settings) for a re-transmission mechanism to handle high rates of PL with much smaller impact on interactivity. However, that low RTT is not the case for today's mobile networks when the service is running at 120 Hz, and thus that exception case can also be ignored.

For UL the effect of packet loss is much more dramatic, a 5% packet loss rate can yield a doubled packet/bit rate. However, the UL traffic is, compared with DL, using much less of the maximum capacity of the link, and therefore UL adaptation for PL can be ignored.

Large spikes in the delay were in the recorded traffic observed to cause a large increase in uplink traffic, possibly warranting a different packet pattern in those cases. Generally though, the traffic amount is still low versus the link capacity and that UL traffic is still, compared with DL, using much less of the maximum capacity of the link. Thus, the UL adaptation for jitter spikes can be ignored.

II.3.2 Adaptation pace

The adaptation of the packet patterns should correspond to the adaptation during the active gaming time (game match) of 90 seconds (value used in the subjective test). A resolution for the measurements on gaming performance over mobile networks within the context of a drive test, needs to ensure a balance between the following:

– drive test real-time characteristic

– feasible (practical) adaptation scheme of the tested service to the network conditions

– expected behavioural changes of the tested service's (aka gaming) interactivity depending on the network conditions

Based on these, a measurement time window of 10 s is reasonable. During the 10 s a 40 ms (or 100 ms as mentioned above) packet pattern is played repeatedly. Within each time window of 10 s, a decision on the adaptation is made after 8–9 s based on the network condition.

II.4 Parameters for approximation subjective scores for cloud gaming

II.4.1 Interactive subjective test

For parametrization of the interactivity model as described in Annex A subjective scores are used in accordance with clause 9.2.1.

The basis of parametrization is an interactive subjective test set-up in a lab-controlled environment. The test used 31 subject gamers (males=29 and females=2) who met the requirement of having experience with playing games, either on computers or smartphones, and of being of age 18 or older. The subjects provided answers to the cloud gaming questionnaire as defined in [b-ITU-T G.1072].

The subjects played a CS_GO FPS game of 90 seconds match rounds. The reason for the game selection relies on the fact that the scope is to provide a generic quality testing solution for one of the most popular video gaming genres played on mobile devices, using technology typical for delivering these, and at the same time addressing the most demanding video game genres for mobile networks in the sense of being most sensitive to the mobile network performance. In this way operators can optimize and troubleshoot their network for the worst scenarios, and thus pre-empt and avoid customer dissatisfaction for both the most demanding as well as the less demanding games.

For each gamer, each of the match rounds have been altered by 30 individual network conditions defined by a set of network metrics. Both the network metrics as well as their values and combinations are selected to meaningfully describe the impact of the mobile network performance on the cloud gaming service based on an interactive subjective pre-test run with a small number of gamers, different from the group used in the full-scale test. The following network metrics are defined as follows:

i) **RTT (delay)** as the median time the system delivers and receives back a packet, referred to as "static" delay.

ii) **Jitter** defined as two types of delay variations which show different effect on the Steam Link streaming.

- ***Random jitter*** frequent with small delay changes (IPDV) from packet to packet.
- ***Jitter spikes*** characterized by amplitude values ≥ 50 ms and possibility of frequency between 0.01–1 Hz.

iii) **Packet loss** as lost packets during transmission or packets which cannot leave the reflecting server due to downlink congestion and being discarded by the server after timeout.

The network conditions have been created using NetEm emulators and a summary is presented in Table II.2.

Table II.2 – Mobile network conditions

Network metric	No. of conditions	Values
RTT	7	2*, 25, 50, 100, 200, 300, 400 ms
PL (per link)	4	0, 5, 25, 45% with RTT=2 ms*
PL (per link) with RTT	9	PL (0.2, 1, 5%), each with RTT/(25, 50, 100 ms)
Jitter spikes (per link)	6	Jitter spikes with amplitudes values of (25,100,750 ms), applied every (5, 15, 45 s)
Random jitter (per link)	4	Avg=25 ms, stdev=3,6,9,12 ms
* An RTT of 2 ms is used as reference for a "clean" condition to be simulated with NetEm.		

The target value used for the function parameterization as in Annex A for the cloud gaming over mobile networks use case is selected to be the answer to the question regarding the overall QoE [b-ITU-T G.1072]. The decision for this approach is based on the outcome of an interactive test for which the results show that the perceived overall QoE describes the interactivity experience as impacted by all the gaming quality dimensions. In addition, this is also empowered by the usage of a specific game with a specific resolution which is acting as a normalization to a reference video-audio quality, and consequently making the overall QoE an interactivity score.

II.4.2 Function parametrization according to Annex A

The principal formula for perceived interactivity is as described in Annex A

$$IntAct = I_{\mathrm{L}} \times D_{\mathrm{PDV}} \times D_{\mathrm{DQ}}$$

The parameter I_{L} is calculated as $I_{\mathrm{L\text{-}CG}}$ by

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

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

The consideration of packet delay jitter is considered by standard deviation if IPDV is used. By standard deviation is meant the average of standard deviation of IPDV on DL and standard deviation of IPDV on UL.

$$\mathrm{IPDV(i) = D(i) - D(i\text{-}1)}, \text{ and}$$

$$D_{\mathrm{IPDV}} = 1 - \frac{IPDV_{\mathrm{STDEV}}}{u}$$

Finally, the effect of packet loss is modified compared to Annex A, because the 3GPP RTT specified limit (RTT = 100 ms) does not reflect when interaction with the service (cloud gaming in this case) is not possible per se, since packets are rarely late for a critical moment, and users adapt to latency which anyway is handled by I_{L}.

The proposed modification is to consider a limit of 100 ms for disqualification on large jitter (IPDV) values in order to capture the fact that large delay spikes disrupt the ability to interact. The reasoning for this is that tests showed that delay spikes start to affect at 50 ms and can have very noticeable effect at 1500 ms, and these jitter spikes are not captured by the median value of PDV.

In addition, it is proposed that DQ term is used to describe the packet loss. Therefore, the D_{DQ} term becomes:

$$D_{\mathrm{DQ-CG}} = 1 - v \times P_{\mathrm{DQ}}$$

with

$$P_{\mathrm{DQ}} = \frac{P_{JS}}{T_{\mathrm{MeasureWindow}}} + \overline{P_{\mathrm{L-CS}} + P_{\mathrm{L-SC}}}$$

and

$$P_{\mathrm{JS}} = \max(\mathrm{IPDV}) \quad \text{, if } \max(\mathrm{IPDV}) >= DQ_{\mathrm{LIMIT}}$$
$$= 0, otherwise$$

where by JS is denoted the jitter spike, and DQ_{LIMIT}.=50 ms

These modification results in a combining formula as

$$IntAct\ for\ CG = f_{\mathrm{offset}} + f_{\max} \times I_{\mathrm{L-CG}} \times D_{\mathrm{IPDV}} \times D_{\mathrm{DQ-CG}}$$

The parameters a, b, u and v to be applied for approximating scores as obtained in the introduced subjective test set-up are:

Parameter	*a*	*b*	*u*	*v*
Value	213	91	25	7

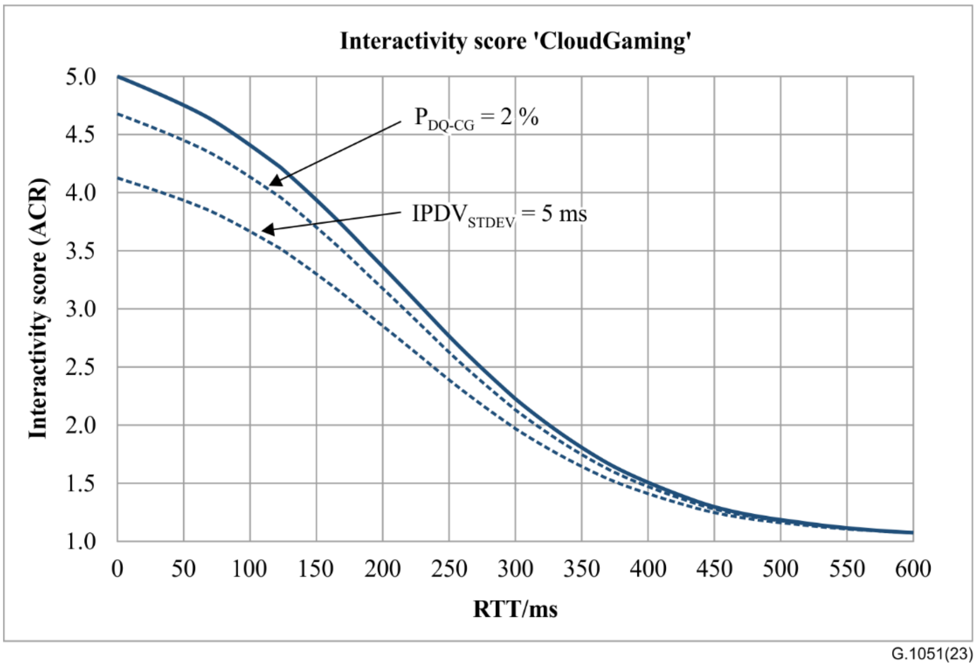


Figure II.1 – Example of interactivity score for 'cloud gaming'

It should be noted that f_{offset} is set to 1.0 and f_{max} to 4.0 to scale the output into a five-point ACR scale as in the underlying subjective test.

Bibliography

[b-ITU-T G.1072]	Recommendation ITU-T G.1072 (2020), *Opinion Model Predicting Gaming Quality Of Experience For Cloud Gaming Services*.
[b-3GPP TS 22.125]	3GPP TS 22.125 (2019), *Unmanned Aerial System (UAS) support in 3GPP*.
[b-IETF]	IETF Draft IPPM Capacity Protocol (2022), *Test protocol for one-way IP capacity measurement*. https://datatracker.ietf.org/doc/html/draft-ietf-ippm-capacity-protocol-04
[b-OB-UDPST]	OB-UDPST (2023), OB-UDPST. https://github.com/BroadbandForum/obudpst

SERIES OF ITU-T RECOMMENDATIONS

Series A	Organization of the work of ITU-T
Series D	Tariff and accounting principles and international telecommunication/ICT economic and policy issues
Series E	Overall network operation, telephone service, service operation and human factors
Series F	Non-telephone telecommunication services
Series G	**Transmission systems and media, digital systems and networks**
Series H	Audiovisual and multimedia systems
Series I	Integrated services digital network
Series J	Cable networks and transmission of television, sound programme and other multimedia signals
Series K	Protection against interference
Series L	Environment and ICTs, climate change, e-waste, energy efficiency; construction, installation and protection of cables and other elements of outside plant
Series M	Telecommunication management, including TMN and network maintenance
Series N	Maintenance: international sound programme and television transmission circuits
Series O	Specifications of measuring equipment
Series P	Telephone transmission quality, telephone installations, local line networks
Series Q	Switching and signalling, and associated measurements and tests
Series R	Telegraph transmission
Series S	Telegraph services terminal equipment
Series T	Terminals for telematic services
Series U	Telegraph switching
Series V	Data communication over the telephone network
Series X	Data networks, open system communications and security
Series Y	Global information infrastructure, Internet protocol aspects, next-generation networks, Internet of Things and smart cities
Series Z	Languages and general software aspects for telecommunication systems

Published in Switzerland
Geneva, 2023