

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

	Page
1 Scope	1
2 References.....	1
3 Definitions	1
4 Abbreviations and acronyms	2
5 Conventions	2
6 Introduction.....	2
7 Test method to obtain two-way latency	2
7.1 Approach to obtain two-way latency	2
7.2 Extending the approach to asymmetrical traffic.....	3
7.3 Guidelines to derive traffic patterns from real applications	4
8 Latency test results and metrics.....	5
8.1 Per-packet two-way latency	5
8.2 Packet delay variation	5
8.3 Inter-packet delay variation	5
8.4 Lost packets	6
9 Interactivity prediction model and related test case definition.....	6
9.1 Principles of the generic interactivity model approach.....	6
9.2 Parametrization of an interactivity model.....	7
9.3 Guidelines for application grouping.....	8
9.4 Guidelines for defining test cases.....	8
Annex A – Computational structure of a generic interactivity model approach.....	9
A.1 Introduction.....	9
A.2 Modelling approach for perceived interactivity of interactive applications...	9
A.3 Consideration of lost and discarded packets and packet delay variation in the model.....	10
A.4 Conclusion	12
Appendix I – Example generic traffic patterns and model parameters according to Annex A	14
I.1 Principle of application emulation and model parameters.....	14
I.2 Examples for application emulation and interactivity score computation.....	14
Appendix II – Parameter implementation according to Annex A based on subjective results for cloud gaming considering a real example application on client and server	20
II.1 Introduction.....	20
II.2 Traffic patterns.....	20
II.3 Guidance for adaptation	21
II.4 Parameters for approximation subjective scores for cloud gaming	22
Bibliography	25



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 packets¹.

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:

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

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.

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.