ETSI TR 103 501 V1.1.1 (2018-10)

Speech and multimedia Transmission Quality (STQ); Guidelines for the Measurement of Data Throughput on Devices connected to Mobile Networks

Reference
DTR/STQ-00217m

Keywords
3G, data, GSM, network, QoS, service

ETSI

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

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

Siret N° 348 623 562 00017 - NAF 742 C
Association à but non lucratif enregistrée à la
Sous-Préfecture de Grasse (06) N° 7803/88

Important notice

The present document can be downloaded from:
http://www.etsi.org/standards-search

The present document may be made available in electronic versions and/or in print. The content of any electronic and/or print versions of the present document shall not be modified without the prior written authorization of ETSI. In case of any existing or perceived difference in contents between such versions and/or in print, the only prevailing document is the print of the Portable Document Format (PDF) version kept on a specific network drive within ETSI Secretariat.

Users of the present document should be aware that the document may be subject to revision or change of status.
Information on the current status of this and other ETSI documents is available at
https://portal.etsi.org/TB/ETSIDeliverableStatus.aspx

If you find errors in the present document, please send your comment to one of the following services:
https://portal.etsi.org/People/CommiteeSupportStaff.aspx

Copyright Notification

No part may be reproduced or utilized in any form or by any means, electronic or mechanical, including photocopying and microfilm except as authorized by written permission of ETSI.
The content of the PDF version shall not be modified without the written authorization of ETSI.
The copyright and the foregoing restriction extend to reproduction in all media.

© ETSI 2018.
All rights reserved.

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

Contents

Intellectual Property Rights

Essential patents

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

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

Trademarks

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

Foreword

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

Throughput, or data rate, is the single most important property of a packet data network. The definition of throughput - transferred volume of data per unit of time - is essentially simple. However, there are many different methodologies available to measure it. To select the most appropriate one for a given purpose, and to assess comparability of results, requires thorough understanding of these methodologies. The present document addresses the measurement methodologies that can be used from an end-user perspective, i.e. embedded in devices or in dedicated test equipments connected to mobile networks.

While there is extensive coverage of IP layer centric methodology (such as ETSI EG 203 165 [i.1] ("Throughput measurement Guideline"), the content of such ETSI Guide does not actually cover methodologies and aspects such as application-level measurements. The present document takes also in consideration methods known as 'crowdsourcing', which have gained considerable audience (and potentially relevance) in the last years but until the time of publication of the present document have not been subject to extensive treatment in the framework of standardization work (however there are activities under way, e.g. the E.MTSM work item in Question 12 of the ITU-T Study Group 12).

Likewise, the present document integrates multi-threaded measurements into a common methodological frame.

The present document has been written to provide a holistic, organized view of the entire measurement process, which also includes elements such as definition of system under test and system boundaries, post processing of data and relation between basic methodologies and their relation to intended targets of measurements.

Modal verbs terminology

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

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

Introduction

Throughput, or data rate, is the single most important characteristic of packet data networks. While its definition - transferred data volume per unit of time - is simple, there is a wide range of possibilities how actual measurements can be carried out. Consequently, it is hard to decide if results of different measurements are comparable of how results from one type of measurement can be used to predict the outcome of another type of usage.

Also, there are entirely different views on performance characteristics of packet data networks. From a low-level perspective (on the IP layer), a network transfers data packets and its performance is characterized by packet transfer time, latency or delay and packet loss rate; also, the variations in time of these metrics, as well as less frequent events such as packet re-ordering have an effect. On higher protocol layers such as TCP, there is no packet loss; lost packages on lower layers translate into lower overall data rates. From an end-user, QoS or QoE perspective, the performance of a network may again vary as the dynamics of a particular application interact with network characteristics and behaviour of other components in sometimes complex ways. In addition, networks typically use resource and performance optimization mechanisms which further increase the complexity of dynamic behaviour.

Mobile connectivity has become an important element of modern life. Both business and consumers have great interest in knowledge about the performance of mobile networks and useful information about this performance is in high demand. There are few actors which have the means to use professional measurement tools to obtain such information. This is one of the reasons why in recent years, a substantial number of companies have emerged which develop and distribute crowdsourcing tools - subsequently termed 'speed test apps' although this description is not entirely correct - to measure mobile network performance. Due to promised or expected cost reduction, even network operators use and rely on such tools today. In some countries, regulators operate crowdsourcing tools too.

From their mode of operation, these tools are considered to effectively measure network performance from an end customer perspective, as the tools run on end-user smartphones. Use cases typically are http or ftp upload and download and therefore results can be attributed to be QoS values and do not represent actual internet speed measurements as understood by e.g. laboratory measurements or assessments in the regulatory context.

Strictly speaking, the scenarios used represent only a small fraction of actual end-user behaviour as pure upload and download only plays a role in use cases such as app download or transfer of larger number of data as e.g. in transfer of photos or videos (typically in e-mail or cloud storage contexts). Nevertheless, such tests play a significant role in the public perception of mobile networks. Through the interplay between public media and PR of mobile network operators, they have a substantial economic impact.

The basic requirement for any meaningful, professional measurement is repeatability. As long as the properties of the network under test and of the test equipment or application stay the same, running the same test is expected to produce the same results. Repeatability allows then comparison. But comparison can also be understood between measurement tools or applications, and in such a case, it requires that the relevant procedures (parameters, set up) of the test are fully documented. This is usually not the case with current 'speed test' applications, and even less so regarding full interoperability, e.g. by open access to servers used as counterpart of throughput testing.

The present document provides a contribution to the evolution of network performance testing towards a professional degree of transparency. This begins with a consistent framework of definitions and technical terms. The elements of the testing process are then described within this context.

Apart from the obvious direct parameters of throughput testing, such as time windows or transferred data volumes, there are numerous other elements which can have an impact on data values obtained. In this sense, methodology and definition of metrics cannot be decoupled from each other. The process starts with selecting the boundaries to the system under test, i.e. insertion or demarcation points. Next comes the way the system under test is accessed. For instance, if the test is run over a radio access network using a mobile device such as a smartphone, the type and degree of influence needs to be assessed. The type of stimulus is likewise important, such as the protocol type, the structure of data traffic (e.g. TCP or UDP based), and the number of parallel connections. Depending on these selections, other choices also become parameters for testing. An example would be to use some kind of real application to create a particular type of traffic, versus using synthetically generated traffic.

The need for careful consideration is not limited to the generating side of measurement data. The way data is processed may also have an impact on results and is therefore subject to documentation and transparency. This applies e.g. to rules about which data to include in computation of results, and which to ignore or discard. For instance, a methodology may require discarding a certain number or range of extreme values to reduce the volatility of results.

Beyond the direct uses of throughput measurement, one of the driving forces is the prospect of using respective data to predict or infer QoS or QoE for a broader range of services. This is of course desirable from a commercial and practical point of view, in order to reduce the complexity of testing as compared to running actual service test use cases. There is no doubt that doing this is possible in principle, using different approaches such as a model-based ones or empirical methods, i.e. a data driven mapping between results from both domains, can be used. It is beyond the scope of the present document to discuss this topic in detail. It is however clear that a meaningful way to do so involves a large degree of caution and professional care - e.g. in calibration and validation of methods. In any case, it will be necessary to gauge the efficiency, data quality the overall effort of such approaches against direct service tests to obtain QoS or QoE parameters by running actual use cases.

1 Scope

The present document provides a systematic overview of methods to measure throughput in mobile networks, with special focus on measurements using a viewpoint at, or close to, application level. Also, it provides a holistic, integrated view of the measurement process, which also includes a selection of methodologies according to intended goals of measurement, and also covers post-processing and data aggregation aspects.

2 References

2.1 Normative references

Normative references are not applicable in the present document.

2.2 Informative references

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

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

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

[i.1] ETSI EG 203 165 (V1.1.1): "Speech and multimedia Transmission Quality (STQ); Throughput Measurement Guidelines".

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

NOTE: The content of this document series has also been used (copied) in Recommendation ITU-T E.804.

[i.3] Recommendation ITU-T Q.3960: "Framework of Internet related performance measurements".

NOTE: Available at http://www.itu.int/rec/T-REC-Q.3960-201607-I/en.

[i.4] ETSI TR 102 678: "Speech and multimedia Transmission Quality (STQ); QoS Parameter Measurements based on fixed Data Transfer Times".

[i.5] ETSI TS 138 521-3: "5G; NR; User Equipment (UE) conformance specification; Radio transmission and reception; Part 3: Range 1 and Range 2 Interworking operation with other radios (3GPP TS 38.521-3)".

[i.6] ETSI TS 138 101-3: "5G; NR; User Equipment (UE) radio transmission and reception; Part 3: Range 1 and Range 2 Interworking operation with other radios (3GPP TS 38.101-3)".

[i.7] Recommendations ITU-T Y.154x series: "Quality of service and network performance".

[i.8] Recommendation ITU-T Y.1545.1: "Framework for monitoring the quality of service of IP network services".

[i.9] IETF RFC 7398: "A Reference Path and Measurement Points for Large-Scale Measurement of Broadband Performance".

3 Definitions of terms and abbreviations

3.1 Terms

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

data path: sequence of entities or elements (physical/virtual) which a data packet transferred between endpoints traverses

endpoint A: local entity in an end to end test scenario (typically, the actual test system)

endpoint B: remote party in an end to end test scenario (in packet data tests, often a server connected to the public internet; can also be a CDN in case of live public content

fixed-size method: throughput measurement method where a fixed amount of data is transferred and the time for transfer is recorded to calculate a throughput value

fixed-time method: throughput measurement method where data transfer is performed for a fixed period of time, and the amount of data is recorded to calculate a throughput value

3.2 Abbreviations

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

API	Application Programming Interface
CDN	Content Delivery Network
DL	DownLoad
FTP	File Transfer Protocol
HTTP	Hyper Text Transfer Protocol
HTTPS	HTTP Secure
HW	HardWare
ICMP	Internet Control Message Protocol
IP	Internet Protocol
IT	Information Technology
ITU-T	International Telecommunication Union - Telecommunication
MDR	Mean Data Rate
NuT	Network under Test
OTT	Over The Top
PoO	Point of Observation
PR	Public Relations
QoE	Quality of Experience
QoS	Quality of Service
QUIC	Quick UDP Internet Connections
RAT	Radio Access Technology
RTT	Round Trip Time
SI	International System of units (Système international d'unités)
TBKPI	Time Based Key Performance Indicator
TCP	Transport Control Protocol
TP	ThroughPut
UDP	User Datagram Protocol
UL	UpLoad

4 Background

There are standards documents that deal, in part quite extensively, with pure throughput measurement at IP level, or with measurements where throughput or round trip time is part of a larger metric for a given use case or service. The main reason for considering round-trop time together with throughput is that TCP throughput is dependent on both the minimum capacity of all path segments, and the round-trip time (because of the feedback loop in TCP flow-control, this would apply to QUIC or any retransmission protocol in general).

The common factor in these standards is that they focus strongly or exclusively on methods based on events of lower protocol levels (typically, the IP or TCP plane). However, the reality at the time of publication of the present document is that there are various tools available - which sometimes appear to be quasi-standards making use events from higher layers - for which there are good reasons, explored further in the course of the present document.

NOTE 1: The term *application plane* is used here as a synonym for Points of Observation (PoO) above the IP layer. Actually, events used in throughput measurement may come from any layer between the API for basic data transfer (e.g. the operating system's socket API) and user-interface indicators in case real smartphone apps are used. According to the principle laid out in ETSI TS 102 250 [i.2], mixing events from different PoO should be avoided whenever reasonably possible.

NOTE 2: Measurement on radio carrier are also possible, but outside of the scope of the present document. For details on this topic, one can refer to ETSI TS 138 521-3 [i.5] and ETSI TS 138 101-3 [i.6].

At the time of writing, there appears to be no document which provides a comprehensive, practically oriented overview of all aspects of the entirety of packet-data network performance measurements, which goes far beyond core measurement methodology and integrates aspects of system boundaries and data aggregation.

While IP-level methods are also mentioned in the present document for completeness, its main focus is on the application plane, and reference is made to documents which treat IP-level measurements extensively, such as ETSI EG 203 165 [i.1] or Recommendation ITU-Ts of the Y.154x series [i.7].

Application-level measurements can be run on practically all devices while low-level data is typically only available if devices are modified, granting apps full system access. However, such modifications, usually called *rooting*, render a device potentially unsafe (typically, this process removes some security features). Even if users would accept that, which is considered very unlikely, the process of rooting is not easy to perform, can lead to permanent damage of devices of something goes wrong, and voids device warranty. Therefore, it is safe to assume that any larger distribution of measurement apps, and in particular crowdsourcing, will have to be based on application-level measurement methods.

This does not mean that application-level methods provide an entirely different point of view. It is assumed that there are clear, deterministic relations between the layers of packet data transfer, so in principle a direct relation to QoS metrics based on low-level events can be established. In this respect, application-level methods provide access to a much broader range of information sources.

Remark: Relations between layers are testable so respective validation can be done when needed.

One aspect has to be taken into account, namely the potential higher degree of device dependency, which requires additional professional care at the level of test design and conduction. Also, application-level measurements may produce less diagnostic depth than measurements on IP level. On the bottom line, the method of choice will be selected based on the type and depth of required information and the relative cost and effort of obtaining it.

5 Basics of throughput measurements

Throughput is defined as transferred data volume per time, or:

$$\mathrm{TP} = \frac{Amount\ of\ Transferred\ Data}{Duration}$$

Where the unit of TP is typically kbit/s.

NOTE: As the amount of transferred data (data volume) is often given in kByte or Mbyte, some caution is indicated. The SI defines prefix k stands as 10^3 and prefix M as 10^6. In the IT world and even in parts of related standardization literature, the prefixes k and M are used for factors based on powers of 2. The prefix k equals 2^{10} (1 024) and M equals 1 024 × 1 024. Obviously, even a single misuse of the k prefix already has a potential error of 2,4 %. This error may multiply for consecutive wrong usages. For technical purposes, the safest way is to not use prefixes at all, but byte values. When this cannot be applied, great care is advisable throughout the documentation chain.

Values for data volume and duration are easy to obtain. However, there are multiple sources for each of them (taken at different points of observation, PoO) and values can only be compared if the PoO are the same.

Practically, events can be closely linked i.e. by a (demonstrable) systematic deterministic relationship (e.g. a function-call chain) so even if the formal PoO is not the same, it can be treated as equal. This applies even to a case where this chain involves a fixed time differential which then would have to be taken into account.

The PoO aspect is not just a formal one, it is also related to the QoS level. If, for instance, a given data transmission path has a non-negligible packet loss rate, the low-level data transmission activity or rate may be high but with no practical use from a user's perspective. Also, transmission protocols may have different amounts of overhead that lead to different user-perception throughput values for the same low-level data rate.

This aspect has a special dimension when using API functions of certain device classes, e.g. the traffic counters of smartphones. When using such data sources, it is necessary to know the level at which these traffic counters are actually counting. At least, the methodology in the context of repeatability of measurement will need to include an element of continuity assurance.

Likewise, the duration needs a proper definition. It is given by two points in time. The throughput of a network under test is not constant over time, in particular at the beginning of a transmission. Therefore, the choice of specific events or otherwise defined points in time will have an impact on the throughput value obtained from a measurement.

In the case of upload data rates, it is important to understand if the data regarded as transmitted is just successfully handed over to the next stage in the chain, or if it is actually transferred across the network under test.

In case of benchmarking measurements, as all channels are using the same architecture and methodology, all aspects mentioned above are less critical. Usually, they do not affect the ranking order of candidates. In case of single-channel measurements, results are usually compared to those measured with other systems or taken at an earlier point in time. Here, it is important to record the used parameters, and to properly understand eventual effects.

Calibration or connection measurements, as used in other fields of engineering, can also be a useful method to ensure consistency. Such measurements would involve measurements on the same system under test using the previous and the new tool or methodology, and producing a statistically relevant number of samples in order to understand eventual differences in output values.

As a summary, throughput measurements are not intrinsically complex. In order to make sure that the results are correct and consistent, there are, however, some principles of proper craftsmanship to be observed.

6 Treating measurement and evaluation methodology as a unit

Practical experience shows that measurement and evaluation methods cannot be treated separately. For a full understanding of applicable evaluation methods and interpretation of results, the method used needs to be known. The single most important example is the selection of file size (or time window) in throughput measurements.

A certain minimum size is necessary to avoid ramp-up and quantization effects. With typical TCP window sizes in the range of 1 Mbit and beyond, and data rates in the order of 100 Mbit/s, a transfer of less than 3 Mbytes to 6 Mbytes will not produce any meaningful results.

Typically, throughput figures obtained from measurements with a particular set of parameters (protocol type e.g. ftp/http; file size) need to be extrapolated to obtain predictions for other parameters, too. It is important to understand how reliable these extrapolations are, considering that a network under test may exhibit a different behaviour if parameters are different. The term *prediction horizon* will be used to elucidate this situation.

For instance, a resource optimization mechanism may decide that FTP traffic has a lower priority than HTTP. A throughput measurement using an FTP download scenario would then not be able to deliver accurate predictions for usage of HTTP download. Likewise, there may be mechanisms such as fair use policies which reduce the effective data rate after a certain amount of data has been transferred, or strategies which do the opposite, i.e. increasing throughput after some time to deliver a better QoE for large-volume transfers.

With upcoming 5G, and the feature of network slicing, it is very likely that the prediction horizon of measurements will be reduced further.

The exact function of current resource optimization mechanisms is not publicly known (and would be different for each network anyway). Also, network behaviour can change. Therefore, it is highly advisable to apply reasonable precautions, e.g. pre-campaign validation runs or result monitoring to make sure that assumptions crucial to the success of a measurement are valid. The safest way is in any case to make sure that test case parameters are sufficiently close to the actual use cases for which the measurement results are intended to be used.

7 System Boundaries

At the core of every network performance measurement, a clear definition of the object under test is required. Figure 7.1 shows the principal components.

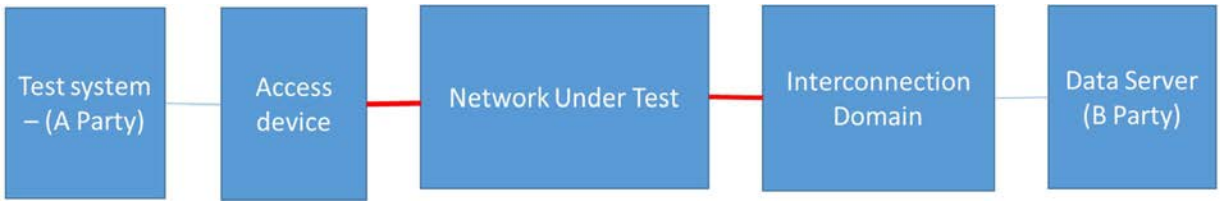

Figure 7.1: Schematic of test environment and network under test

The following practical terms are introduced:

- Endpoints: The principal origin and destination element/device of data packets used for performance testing. In some contexts, the terms 'A Party' and 'B Party' are also used.
- Data Path: This is the route for the data packets which are transferred between the endpoints.

Figure 7.1 describes a functional architecture, with the main purpose of defining the boundaries between the Network Under test (NUT) and the surrounding elements.

In an actual test setup, some of the elements may be combined in one physical unit (as in the case of a smartphone where the A Party test system and the access device are integrated). Likewise, a functional unit may be distributed in physical entities, or the B Party might not be a single data server, but a content delivery network made up of thousands of physical or virtual servers in different locations, with some load- or latency-optimizing logic on top.

The 'Interconnection Domain' element has a special meaning. Even if the purpose of a measurement is to characterize a particular network, its principal connections, at the exact edge of the network to the outer world, may not be accessible. In this case, suitable access or interconnection points are used that constitute part of the data path. This is more than a technical issue. Network performance metrics are often a part of regulation. A network operator can only take responsibility for properties which are under his control. Therefore, the definition of system boundaries is a sensitive issue, which requires great care.

For further reading on this topic, it is recommended to refer to Recommendations ITU-T Q.3960 [i.3], ITU-T Y.1545.1 [i.8] and IETF RFC 7398 [i.9].

For QoS measurements, the endpoint B is often some kind of public server, in particular for web browsing tests which at the time of writing use a mix of some static reference pages, and some live web pages. For performance measurements, a similar architecture with some general counterpart servers is conceivable. Such architectures require, however, some kind of additional validation structures or processes to ensure that third-party inflicted effects do not compromise measurements. Therefore, performance measurements typically use dedicated servers so that both endpoints of performance measurements are under the control of the entity performing the tests.

8 Points of control and observation

Any performance test requires a minimum set of information, which is typically comprised of events on some protocol layer. On lower layers (e.g. IP), the sources of such events are typically traces, i.e. software functions delivering information on packet level. For application-level testing, activities are started, and events are observed, using primary functionality provided by the operating system (API, application programming interfaces), or corresponding interfaces of other software packages on top of such APIs.

Controlling and observing activities can use the same or different interfaces. For instance, an active throughput measurement needs to send some data, which is usually on the API level or higher. For IP level based measurement, this activity is then monitored at another place in the data chain. For application-level testing, the location for control and observation can be the same.

A special case (and one of the weak points of IP-level testing) is a situation where data transfer is encrypted (e.g. using HTTPS instead of HTTP) or uses a proprietary protocol (in case of testing using OTT functionality). This renders low-level events completely inaccessible, or, at best, imposes additional processing requirements such as some kind of pattern recognition to link low-level to application-level activities. Such features are also often device-dependent.

9 Measurement equipment considerations

There are many ways to realize measurement equipment. As long as some basic common sense requirements are fulfilled, there is no better or worse equipment. Even if there are more or less precise instruments (ranging from a microsecond instead of millisecond time resolution to extensive hardware-heavy solutions such as temperature stabilized modems or smartphones), their practical use needs to be evaluated against the practical advantage on a cost to value axis, on the background that in a live network properties fluctuate on significant scales.

There are, however, some basic properties, which need to be fulfilled in any case to make a measurement valid. The most important of these properties is that the measurement system will not have uncontrolled effects on the measurement values.

With respect to throughput measurement, the strict version of this requirement would be that the system may not be the bottleneck in a measurement, i.e. limit by itself some of the values it measures. This demand, however, may be too strict in an efficiency-oriented frame. For example, it may make sense to use bandwidth-limited servers or systems. These would of course not be adequate for measuring the peak data rate a NUT can achieve. It may still be a good choice if the primary interest of a measurement is to check if a network can provide a given minimum data rate. The advantage of a data rate limitation, on the other hand, is that the amount of data transferred is kept in check. The practical value, e.g. in cases where a high speed data volume limit exists for standard SIMs, can be higher than the potential drawback of not knowing the peak data rate of the NUT.

Well-understood restrictions may also be related to certain operation parameter ranges. For instance, running IP Trace imposes - at contemporary data rates of several 100 Mbit/s - considerable extra load to a system in order to process or even just store a huge amount of extra data volume. Assuming that basic measurement system performance is adequate, there still may be other effects, such as a drift in time stamps of captured packets for transactions longer than a given duration, due to a data queue which runs up when the system's mass-storage limits the speed.

The essential requirement is therefore that the relevant properties of the testing system will be well characterized. It will be possible to know if a particular measurement value is a property of the NUT, or somehow influenced by the system.

A special case to be considered, in the context of comparability of measurements, is continuity. Realistically, any measurement system needs to be kept up to date, which involves hardware or software updates or upgrades. For instance, given the speed of evolution in RAT, a given modem or smartphone will have to be replaced, sooner or later, with a newer and usually more capable model. Also, operating system updates will have to be performed, if for no other reason than to keep up with typical user perspective. Another important element of consideration for comparability of measurement is the possibility to run it across several radio access technologies (3G, 4G, WiFi, etc.), including situations where hand overs between these technologies occur. If comparability in such situations cannot be ascertained, then measurement systems needs to be able to block RAT, frequency, and in any case to report under which technology a given measurement has been performed.

It is desirable that a system upgraded in one of these ways delivers exactly the same values as before. Practical experience shows, however, that there are cases where the effort to achieve this goal is unreasonably high. A solution can be using connecting measurements, a technique which is used in other fields of metrology. This technique is useful for both validation of a new system based on a previously existing one, or to establish mapping relations between results of the previous and the new system.

10 Measurement Modes

10.1 Background

10.1.1 General

A throughput measurement can be done by transferring a fixed amount of data (fixed-size method) and measuring the time required, or by transferring data for a fixed amount of time (fixed-time method), and recording the amount of data transferred.

NOTE: Several names or terms have been used for the fixed-time and fixed-size methods. The fixed-size method has been called "best effort method" in ETSI EG 203 165 [i.1]. For the fixed-time methods, terms such as "windowed (ETSI EG 203 165)", "Fixed Data Transfer Time QoS (FDTT-QoS)" (ETSI TR 102 678 [i.4]), "time based", or TBKPI are being used - this list is not even intended to be complete. The choice of the name pair fixed size/fixed time used in the present document is intended to clean up this situation with a consistent and self-explanatory pair of terms.

Both methods have their specific properties, benefits, and drawbacks.

10.1.2 Fixed-size-method

The fixed-size method is, on first glance, more user experience related: the typical use case is a transfer of some kind of data file. For measurement purposes, it has, however, the disadvantage that it introduces a large variation of time per transaction. The spread of data rates in the field current ranges from less than 100 kbit/s for 2G up to 300 Mbit/s for 4G. A long transaction time is undesirable in several ways. In benchmarking, it inevitably leads to a very fast de-synchronization of transaction types in mixed scenarios. At best, with pure upload or download scenarios, the effect is a large difference in sample count which is also undesirable. Last but not least, in drive test scenarios long transaction times lead to small sample densities or low spatial resolution.

To end up with a reasonable maximum transfer time for 2G, file sizes would have to be so small that the peak data rate in 4G would not be visible due to ramp-up. The only way left to at least dampen this effect is a time-out which cuts off a transaction after a given maximum time. Applying the usual validity rules for samples in QoS measurement, the downside of this solution is that affected transactions would be removed entirely from evaluation which makes the whole effort pointless.

Adaptive solutions have been proposed, which select the measurement parameters according to the RAT shortly before the transaction begins. However, this solution does not work well as RAT selection is part of mobility control and therefore the situation typically changes frequently.

In conclusion, fixed-size methods work well, at best, in very static environments such as laboratory tests or drive tests in RAT-wise homogenous networks.

10.1.3 Fixed-time-method

The fixed-time method has none of the disadvantages of the fixed-size method. The only drawback is that for low-throughput RAT, the user perspective is not given. Theoretically, a network could regularly (or with a given probability) fail to transfer files above a certain size. This would go undetected if the data volume transferred within the time window stays below that value. However, in the end a decision for one of these methods needs to be made. Additional tests to uncover such blind spots could be added if there is evidence that they actually occur.

There is, however, a situation where the fixed-time method may deviate from a strict timing pattern. This occurs when the data connectivity cannot be established, or is interrupted during the transfer. This would represent, as for the fixed-size method, a transaction result of 'failed' or 'dropped', respectively. If maintaining a constant time pattern is required, mechanisms to insert an extra pause are required.

However, extra consideration is advised in the case of extra pause. In drive-test situations, every pause means a gap in spatial information. If a vehicle moves at a speed of 50 km/h, an extra pause of 5 s equals a blind area of 70 m on top of what already occurs during regular pauses. The argument that the higher density of values in the case of unsuccessful outcomes creates a bias towards a more negative assessment of network performance, is, of course, true. On the other hand, given typical values of residual pauses, the effect is moderate and provides the benefit of a higher spatial resolution of problematic points in the network. It could even be argued that a slight over-emphasis on poor-performing parts of the network is beneficial in the sense of achieving a high network quality.

Also, in the case of dropped or failed transactions, it is necessary to define if or how respective data are to be processed for aggregation. For an average throughput, the choice is between ignoring these values, and using a functionally equivalent value. In the case of a failed transaction, the resulting effective throughput would be zero as the user could not transfer any data during this time. For an interrupted transfer, the throughput could be defined based on the amount of data transferred before the interruption, divided by the whole length of the time window.

10.4 Selection of the most appropriate mode

From a technical point of view, fixed-size and fixed-time methods are actually quite similar. A fixed-time measurement can be understood as a data transfer with an infinite file size and a time out at the value of the intended time window. With the fixed-size method, a transaction which is stopped by time-out is computed as unsuccessful. So, the only element which has to be modified for processing fixed-time measurements is that a timed-out transaction is counted as a successful one, i.e. its data being processed for averaging.

When processing fixed-time measurements, some attention has to be paid to the initial part of the transaction, which is characterized by an increase of throughput towards a saturation value (ramp-up phase, as visualized in Figure 11.1). In order to reach this saturation zone, the duration has to be considerably larger than the time of throughput ramp-up towards its stationary or saturation value. Practical time window values are then in the order of 10 s.

This poses another issue, as in RAT that provide high data rates, such as high-throughput 3G varieties and 4G, the amount of data can easily reach several tens of MBytes. This is usually much more than the data volume necessary to reach the saturation throughput in such fast RAT. Especially when using "commercial" SIMs with typically limited "high speed" data volume, this leads to unnecessary depletion of this volume. But even for unlimited data volume, transferring such large volumes puts an unnecessary stress on the network. In addition, the amount of supplementary measurement data (typically mostly from IP Trace) and related effort to store and transfer this data scales accordingly.

Therefore, a third alternative to fixed-size and fixed-time methods is a combination of both, which is termed 'hybrid method'. The basic method is fixed-time, but with a finite file size. This size will be large enough to reach the saturation throughput, but usually can be dimensioned such that in fast RAT, the amount of data transferred is considerably smaller than it would be for a fully used time window, whereas still sufficient to reveal the NUT's throughput capabilities.

If the size limit is reached, the transaction ends. That means that a variable extra pause would be required to keep a constant time frame. Similar to the case of failed or dropped transactions discussed above, it is a matter of priorities: a larger effective pause means a lower spatial resolution.

Under practical aspects, the biggest advantage of the fixed-time method is the almost constant temporal rate of sample points it produces.

In case of total loss of connectivity, the frequency of sample points is still not constant. The desirability of artificially extended pauses is a matter which needs to be carefully assessed.

Under drive-test conditions, this also translates to a constant spatial resolution of measurements. Furthermore, when transaction-wise sample points are aggregated towards an average data rate, the fixed-file method would create a systematic bias towards a higher mean data rate due to the relatively higher number of sample points from areas with higher throughput.

A possible aspect in favour of the fixed-size method would be its closer relation to user experience, as most practical use cases deal with the transfer of a fixed amount of data. It can be assumed that all contemporary mobile networks use, at least to a certain degree, resource optimization methods which in effect adapt the network performance to the given use case (content awareness). Modelling the scenarios used for testing as closely as possible to the situations for which a QoS assessment is made is therefore a matter of securing the validity of such an assessment.

Mitigating the effects of the fixed-size method on session times and sample count is possible if time-outs are used. However, the side effects of such parameter choices have to be considered thoroughly. Mainly, this is the filtering effect on values used for statistics (such as averaging). The standard processing mode is to use only throughput values from successful transaction for statistics. A time-out that limits the maximum session time can therefore have a systematic biasing effect by removing many values from the lower end of the throughput spectrum from an average. Therefore, when using this method, it will be necessary to review and eventually adapt processing accordingly.

When measurements are to be interpreted from a QoE perspective, reporting session times may be preferable to reporting throughputs, for instance for downloading or uploading a data volume typical for a given group of network users. Such times can be calculated from throughput values in a straightforward way. It should, however, be kept in mind that set-up and network resource ramp-up times are a reality in actual user experience. Calculation of realistic session times will therefore need consideration of such elements, as well as take into account how the throughput values have been computed in the first place, that is, if they represent upper-limit values or if they already factor in ramp-up times.

The actual choice of the method and its parameters used in a particular test should therefore reflect the goals of the measurement and the target audience, and also include a careful consideration of the network performance characteristics to be expected.

10.5 Practical examples

To get a numerical impression of value ranges, see the following tables.

NOTE: The values shown are not the maximum values for respective technologies. They can be understood as typical values in a practical, order of magnitude sense.

Table 10.1: Numerical examples for download indicators vs. RAT

	2G	3G	4G
Typical DL TP [Mbit/s]	0,1	20	100
Typical DL Transfer time 1 MByte [s]	80	0,4	0,1
Typical DL Transfer time 10 MByte [s]	840	4,2	0,8
Typical DL volume in 10 s intervals [Mbytes]	0,13	25	125
Typical DL volume in 3 s interval [Mbytes]	0,04	8	38

Table 10.2: Numerical examples for upload indicators vs. RAT

	2G	3G	4G
Typical UL TP Mbit/s	0,05	5	20
Typical UL Transfer time 1 MByte [s]	170	1,7	0,4
Typical UL Transfer time 10 MByte [s]	1680	16,8	4,2
Typical UL volume in 10 s interval [Mbytes]	0,06	6	25
Typical UL volume in 3 s interval [Mbytes]	0,02	2	8

Table 10.3 shows some per-RAT examples for expected RTT ranges, in conjunction with ramp-up times. As the TCP slow start is not the only factor which determines ramp-up, an additional "ramp-up time frame" value is used. The estimated ramp-up time (to approximately 95 % of the saturation value for throughput) is calculated as min(5 × RTT, Ramp-up time frame).

Table 10.3: Numerical examples for RTT and time to throughput saturation vs. RAT

	2G	3G	4G
RTT factor to saturation	5	5	5
Ramp-up time frame	3	1	1
Typical RTT [ms]	800	400	100
Assumed ramp-up time to TP saturation [s]	4	2	1
Typical ICMP ping time [ms]	220	70	50

11 Data Evaluation

11.1 Basic considerations

A throughput measurement will usually create a sequence of data points where each data point, at a given time t(N), represents the data volume transferred between this point in time and the previous one, t(N-1). Depending on implementation, data volume values come directly from respective points of observation, or are calculated as differentials from platform-specific information sources such as traffic counters.

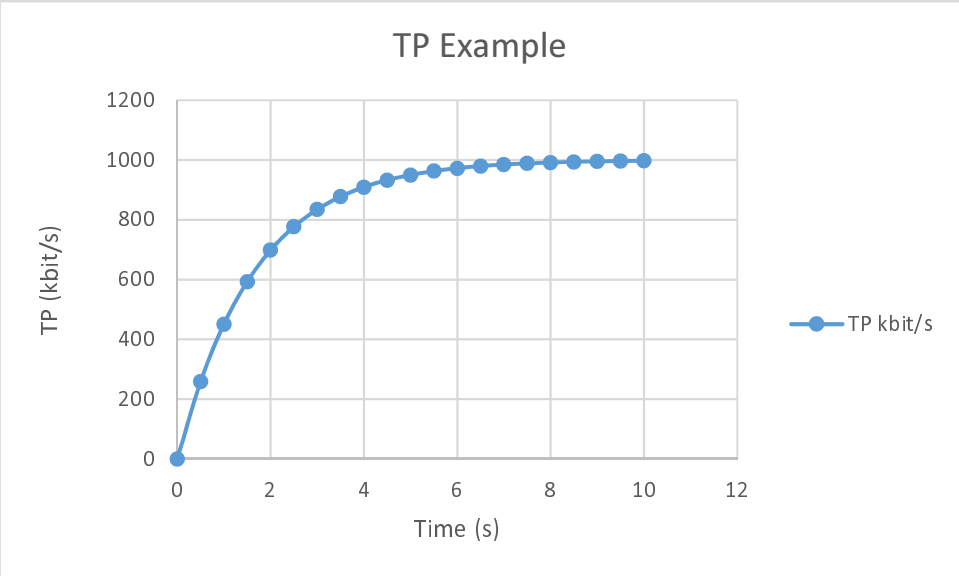

Figure 11.1: Example for a throughput vs. time diagram

Remark: In some cases, and depending on the actual use case and the points of observation, there may be just a single data point. It may also be that the data volume is not actually measured, if the use case is the transfer of a data object of known size.

In practice, throughput is not constant over time; typically, there is an initial ramp-up phase. Figure 11.2 shows an idealized example. In reality, the shape will typically have a more complex structure, where packet transport layer behaviour such as TCP Slow Start, but also network resource management related behaviour plays a role.

The throughput for the whole transaction will be calculated from these data points. In the simplest case, the time window used is just the difference between the last and the first timestamp of the sequence, and the transferred bytes are calculated by summing up values from respective data points.

This is - with respective time window start and end definitions - identical with the Mean Data Rate as defined in ETSI TS 102 250-2 [i.2].

11.2 Test case parametrization and post processing aspects

By selection of reference points in time, it is decided how much of the ramp-up phase is taken into account. Figure 11.2 shows an example with t1 and t2 being the chosen start and end points in time.

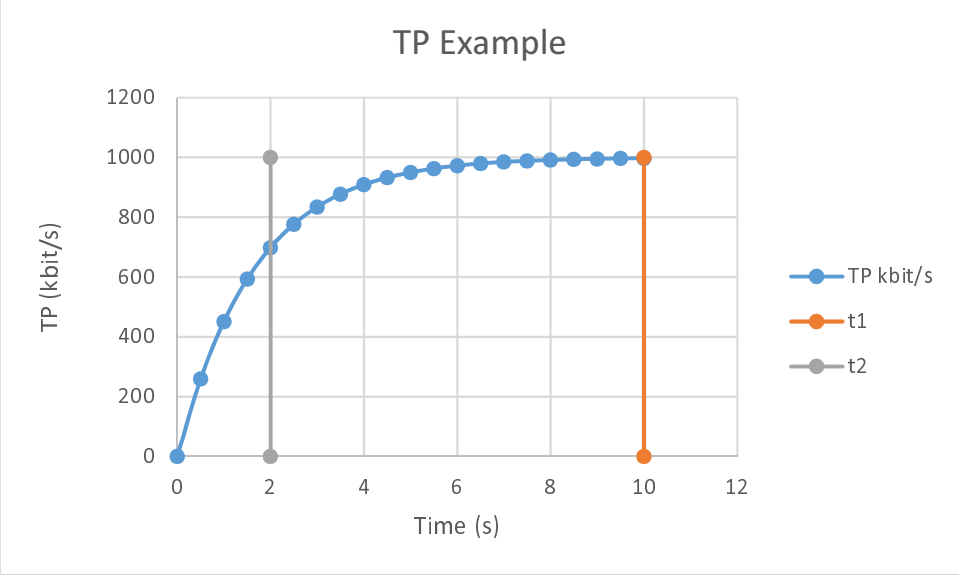

Figure 11.2: Typical throughput vs time diagram with start and end time window markers

The selection of t1 and t2 depends on the purpose or objective of the measurement. From an end customer perspective, the ramp-up phase is part of the user experience. Also, the data volume transferred, or rather its order of magnitude, relates, to use cases. Therefore, t2 and, subsequently, t1 would be selected to relate to typical use cases if the purpose of a measurement is dominated by this aspect.

If, on the other hand, the objective of a measurement is to produce information about the maximum data rate the network can deliver, the parametrization of the test case would need to ensure that the saturation throughput is reached (sufficiently large t2) and t1 would then be chosen to exclude, or largely exclude, the ramp-up phase. With sufficiently large t2, the impact of ramp-up becomes smaller in any case, i.e. the choice of t1 becomes less critical.

Selection of t1 and t2 are made in post-processing, i.e. a given set of measurement data can be evaluated from different angles. The available range is, however, given by the parameters used in the measurement which usually considers other aspects, such as the amount of data transferred per transaction, and the possible side effects of intensive resource usage on other users at the time of measurement.

Remark: Apart from activities of other users in the network, it has also to be taken into account that background activities on the device used for measurement can take place. As far as possible, set-up needs to prevent this. Where this is not possible, as e.g. in crowdsourcing applications, respective measurement data need to be taken in order to consider these effects in post processing.

Considering data volume aspects is especially important in crowdsourcing-type measurements where using up subscriber's data volume may have an impact on the general motivation of end users to participate in crowdsourcing in the first place. Furthermore, the question is not only the volume of data transferred at a given time, it is also about the frequency and time of measurements (if this occurs every 10 minutes in peak hours of only once a day at 3 am, the impact will not be the same, etc.).

Actual time sequences of throughput measurements will exhibit fluctuations which can be very strong, depending on time resolution and circumstances of a measurement. Figure 11.3 shows a (constructed) example of a noisy data sequence. The same generating function was used with a random component added (the connecting lines are for better visibility only).

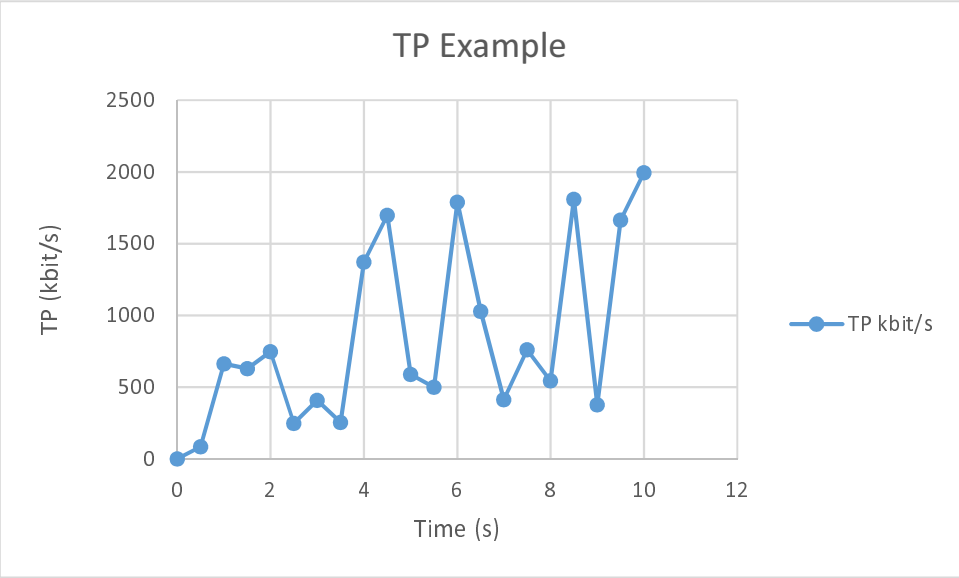

Figure 11.3: Throughput time sequence with noisy data

The sources of fluctuations are manifold; a detailed discussion is not within the scope of the present document. There is, however one aspect related to the packet nature of data in connection with limited time resolution. On the IP level, packet size is typically around 1 500 Bytes, but on the application level, it may be considerably larger, depending on IP stack parametrization. The effects also depend on the method of time stamping. If fixed time intervals are used for sampling transferred data volume, a small variation in arrival time can lead to a large variation in data volume for the respective time interval. It is advisable to select a time scale where typically a larger number of data packets arrive to sufficiently dampen this effect. If timestamping based on packet-arrival events is used, it is advisable to make sure the typical frequency of events is large enough to keep quantization noise at a reasonable level.

As far as the fluctuations can be considered effectively random, increasing the time window t2-t1 reduces fluctuations in any case. Typical time windows used in throughput measurements towards Mean Data Rate values are ranging from 5 to 10 seconds upwards. For analysis purposes, typically time resolutions in the order of 0,5 to 2 seconds are used.

11.3 Using subsets of data points

Processing raw data, i.e. time sequences, from throughput measurement case use more elaborate algorithms than just aggregation of data within a given time window as described above. Again, depending on the purpose and objective of a test, it may also be useful to further select data points, i.e. include or exclude such points or whole sections of a sequence. Figure 11.4 shows an example to explain this further.

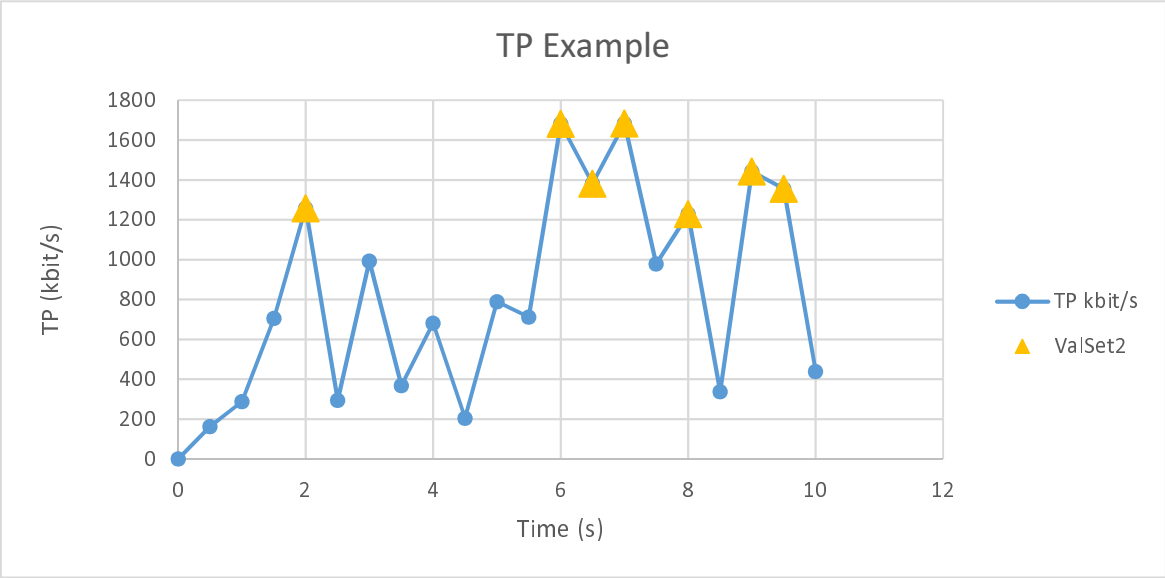

Figure 11.4: Example for selection of values from a throughput measurement.

One of the simplest ways of data point selection is to remove extreme values (the highest and/or lowest M values of the set) to reduce variations from subsequent measurements.

If the purpose of the measurement is to provide information about the maximum performance the network under test can deliver, selection of the K best values (or, in the extreme, only the single best value) can be appropriate. Another variant of such a method would be to use a moving window over the whole data set and use the highest value obtained.

If it is desirable to limit the amount of transferred data, as previously mentioned e.g. in crowdsourcing applications, another way to process data would be to use a rather small time window and try to estimate the maximum throughput by extrapolation from the ramp-up phase. Another possibility is a careful sampling of measurement points as far as this is reasonable with respect to quantization noise and processor load. Such methods may be refined by using other information, e.g. from lower protocol layers.

The applicability of any particular method is not a matter of absolute judgment or restrictions. The assessment of their usefulness with respect to accuracy, reliability and value vs. effort is a case by case matter. In a professional context, it will however be important to provide transparency and repeatability. This means that the methods used to compute data towards reported values are part of the information given about the measurement.

11.4 General aspects on reporting of throughput measurement results

Apart from direct technical aspects of throughput measurement, there may be another group of aspects which relate to general QoS and which may be relevant for consideration in a given measurement or reporting situation.

These aspects deal with the actual applicability or results to predict end user experience. A given mobile network may be technically able to deliver a certain performance with a particular kind of access (represented by the SIMs which are used by the measurement equipment) and/or a given testing method, e.g. a multi-threaded download.

The first aspect is the question of availability of subscriptions or data plans to the general public. The information about the performance of a mobile network is only relevant for a typical end user if a product can be purchased.

Remark: There are special cases, such as lab tests or trials, where this condition is not applicable.

The second aspect is about the applicability of measurements done with a particular use case, e.g. an FTP download, to a situation where another protocol, e.g. http, is used.

The third aspect relates to any traffic shaping, e.g. resource balancing policies. For instance, throughput may firstly increase over time until a saturation or stationary value is reached, but then be reduced again after a certain data volume or time has been reached. Likewise, throughput may reach a first saturation value, but after some time the network may allocate additional resources to provide a further increase in throughput.

Last but not least, assessment of the circumstances of a measurement may be necessary. For instance, a measurement may be performed at low-load hours to minimize negative effects for other users. In that case, results may not represent the situation experienced by a typical user under normal or peak network load conditions. Consideration of the outer conditions of such measurements is therefore also a factor which may be part of the design of a measurement, or part of the documentation of such measurements.

12 Considering equipment related effects

A standard safety control question for throughput measurements is if there are systematic effects which can affect measurement results. The most serious effect would be an element in the access part (i.e. in the test system domain) which limits throughput.

For TCP based packet data transfer, there is an upper limit to the data rate given by IP stack settings. While it may be assumed that the default settings of smartphones do not limit the data rate, this assumption needs to be validated before a measurement system is actually used. This can be done by a combination of the following:

- Query the appropriate API functions provided by the operating system;
- Use IP traces to obtain that information;
- Run actual validation tests to make sure that the device's IP stack is not the data rate limiting element in the chain.

In typical smartphones, another effect has to be considered. There will always be a certain amount of parallel traffic, i.e. data transfer by other software running in the background while the throughput tests are being performed. Large-volume data traffic - such as background downloads of app updates - needs to be suppressed anyway by an appropriate set-up of the device. However, experience shows that such traffic cannot be completely suppressed. It is strongly advised to check the global byte counters to make sure that the amount of such background traffic is reasonably small compared to the target traffic.

13 Latency measurements

A performance test usually also includes latency measurements. Traditionally, this is often done using the ICMP *ping* function. There are some reasons to question this tradition, and to consider alternatives:

- ICMP *ping* is a low-level function. It and can be understood to provide a lower estimate of a practical round-trip time rather than a time which can be expected for actual packet base message transfer between two entities on the user plane.
- Some networks or servers suppress ping as part of a protection strategy.
- On some platforms, use of ping (by some operating system related shell functionality) might be restricted, e.g. requiring rights above user level (root access).

Alternatives to be considered include the use of suitable transfers on higher layers such as HTTP or FTP. In this case, there is a small extra delay due to reaction times of server processes. On the plus side, an RTT measured this way may give a more realistic picture of latency in user-related scenarios (e.g. gaming), and increase robustness of the measurement against selective suppression of traffic by elements in the data path.

Such solutions can even provide a more efficient solution, where no extra time is needed for RTT measurement, by way of using parts of sequences for other parts of the performance test.

Other than that, an RTT measurement is rather straightforward. It produces values which can easily be further processed by applying standard statistical rules with respect to accuracy.

The constraints of latency measurement are equivalent to those for throughput measurement. End-to-end latency is composed of contributions from the network elements, protocol stacks at receiving side, application, etc. A good knowledge of PoO is also required, as well as a distinction amongst these various sources of extra delay (and hence responsibilities).

14 Aggregation

14.1 Overview

The general term aggregation includes different ways of computation to condense individual data items into larger units. Aggregation per se is rather simple. With aggregation, statistics come into play, which means that an aggregated value (as typically in QoS matters) is also a prediction how a user will experience the system under test under the same or similar circumstances. Mainly but not only for this reason it is important to understand what the actual meaning of a value created by aggregation is.

14.2 Temporal or data-point aggregation

The goal of this kind of aggregation is to create a single data point from a couple of input data points.

Often, the term *average* is used to mean the arithmetic mean. It is however important to notice that there are other types of averaging, such as median values. Before this aspect is treated in more detail, there are some general remarks to be made about the basics of processing, in particular with respect to throughput measurements.

A throughput measurement usually outputs time-resolved values. Measurements using points of observation at lower-levels (IP or TCP), when looking at individual packets, can produce time resolutions in the millisecond range. Data noise is high, at least in field measurements as compared to laboratory measurements, so usually a primary aggregation to time scales of some 100 ms is used. With points of observation at higher protocol levels, time resolutions are typically also in or above this order of magnitude. Such information can either be an event such as the arrival of a data packet of certain size, or is polled by e.g. querying some type of traffic counter available in the system (see also the examples shown in clause 11).

The polling method - assuming that its frequency is chosen to not have negative side effects due to overload to system resources - may be preferred because it creates a more reliable time pattern, which is particularly helpful in periods with low or zero throughput.

There are situations where it makes sense to use this time resolution to create a spatial resolution, as in drive tests where spatial aggregation can be used to reduce data noise (see also next clause). One way or another, aggregation of throughput values will be required, however, to create a respective QoS parameter.

There are two possible ways to average these values:

a) Compute the average of individual throughput values (from time interval and amount of data transferred in this period of time);

b) Compute the throughput as the sum of all data transferred divided by the sum of all intervals.

Depending on the distribution of values, the results of these two methods can be very different. In method a), the result is dominated by high values. Unless such high values are not artefacts (e.g. by extremely small time intervals and residual packet-related granularity), this contains information about the network capability to deliver such high throughput. In a typical user scenario, a short peak in performance alone does not have a practical value if the duration, or the transferred data volume, of a typical transaction is much larger than the corresponding samples. Therefore, using method b) will usually be the method of choice when a measurement is aimed at getting information related to user experience.

For creation of indicators covering a number of subsequent transactions, basically the same considerations apply. In that case, however, the sequence of transaction-wise values represents repeated probing of the same system. To select the best fitting computational method, the appropriate type of averaging also needs to be taken into account.

Practically the property of method a), which tends to overemphasize large individual values, needs consideration from a different angle of view. When the average of a series of measurements aims to produce a prediction of the most likely value, then the median value may be a better choice than the arithmetic mean.

14.3 Spatial aggregation

A network performance measurement can be made stationary, i.e. in a fixed location, or it may include motion through a given geographical area (or a laboratory situation with a simulated radio access environment).

Again, the way data is aggregated needs to be considered against the purpose of the test, or the actual meaning of the numbers created by aggregation. Here, another dimension comes into play when the network under test is a typical mobile network with mixed RAT.

Aggregation creates, in the form of metrics, a verdict for the whole area. Using averaging implies, strictly speaking, that the domain which produces the measurement values is homogenous and fluctuations in values are random. However, it is clear that this is not the case in a region where different RAT (or different variants of the same RAT or different parametrizations of access points) are used.

Consider the following example:

- There is a region with ten spatial segments. In one of them, 4G is deployed with a DL throughput of 100 Mbit/s. In the others, there is EDGE with 200 kbit/s. Now a drive test is made, with some pre-processing yielding one value per segment. As long as the number of samples per segment is the same, the example works just the same for other groupings of data.
- The arithmetic mean of these values produces a value close to 10 Mbit/s as the single high value completely dominates this type of average. Even method b), calculating throughput as quotient of the sum of all data volume, and sum of all times, would yield a similar value.
- To understand the resulting value as a prediction, there are two possibilities. If the use case describes a single transaction with sufficiently fast motion across segments, or the average of a series of transactions, the measurement results would have predicted the experience correctly. If, however, the user stays within a segment for the time of a transaction, 9 out of 10 cases the throughput actually experienced would be far below the predicted value. The situation becomes extreme when the motion is not part of the target scenario, but a means to change locations. A stationary user would, in 90 % of the area of this region, experience a throughput which will never come even close to the measured value.
- These considerations do not mean that aggregation should not be done. They are, however, reminders that the type of aggregation applied should be carefully checked with respect to the meaning carried by the results. The average is not necessarily the only way to aggregate data, other types of statistics can also be imagined in complement.

15 Multi-socket measurements

While there is no explicit reference to the number of parallel activities used, method and QoS parameter definitions in ETSI TS 102 250 [i.2] clearly use the assumption that only a single IP connection (socket) is used. In networks existing at the time of writing, full specified data rates are usually achievable only if more than one socket is used. Therefore, the goal is to extend the framework for throughput measurements in a generic way. This means that the number of sockets becomes just another parameter for measurement, and that clear rules exist with respect to definition of valid samples, start and end of transactions.

NOTE: This clause considers multi-socket scenarios which use the standard IP connectivity available through the mobile device's respective API. It is therefore the equivalent of end user perspective on application level and does not deal with possible effects of differentiated QoS applied to radio bearers. In order to explain the concept in a meaningful and consistent way, the following set of working assumptions and definitions is made. To be clearly understood, this does not claim to be normative in any way, and other sets of assumptions are not excluded:

- A single instance (*transaction*) of a multi-socket test case consists of N *actions* which are started in parallel:
  - The type/composition of the test case is a parameter. E.g. 3x HTTP, 4x HTTP, or 1x HTTP and 2x FTP are different tests which need not to be mixed in evaluation.

- A validity rule for a transaction defines which transactions are included in result computation:
  - The core time window of a multi-socket transaction starts when all sockets are transferring data.
- If not all designated transfers have been started (within a given time window), the attempt is rated 'failed'.
- The criterion is that at least one packet of content is received on each socket.
- Remark: There may be situations - e.g. when using an app which supports a variable number of sockets and no in-depth information such as IP Trace is available - where it may not be possible to determine this condition. Whether this is tolerable or if further requirements to the testing system would have to be added, will depend on the specific diagnostic goal.
- The transaction ends when the first socket has transferred all of the data assigned to it, or the end of the transfer time window is reached:
  - If one of the sockets drops, the transaction result is 'dropped'. Otherwise, the transaction is considered to have ended successfully.

Figure 15.1 and Figure 15.2 illustrate these assumptions graphically. Figure 15.1 depicts the ideal situation with synchronous start and fully balanced performance of all sockets. Figure 15.2 shows the realistic case where transfer starts at different points in time and data rates are different between sockets.

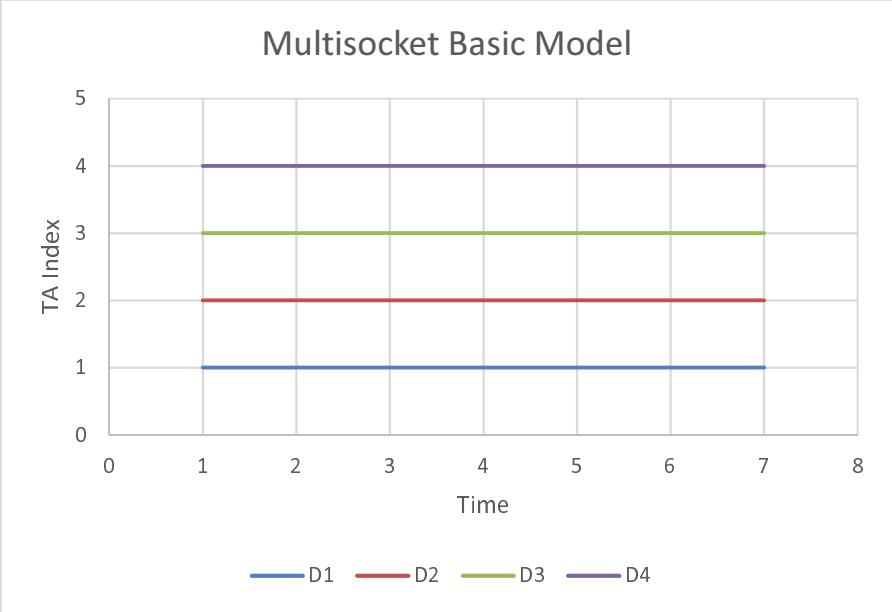

Figure 15.1: Basic ideal multi-socket model where all transfers start and end at the same time

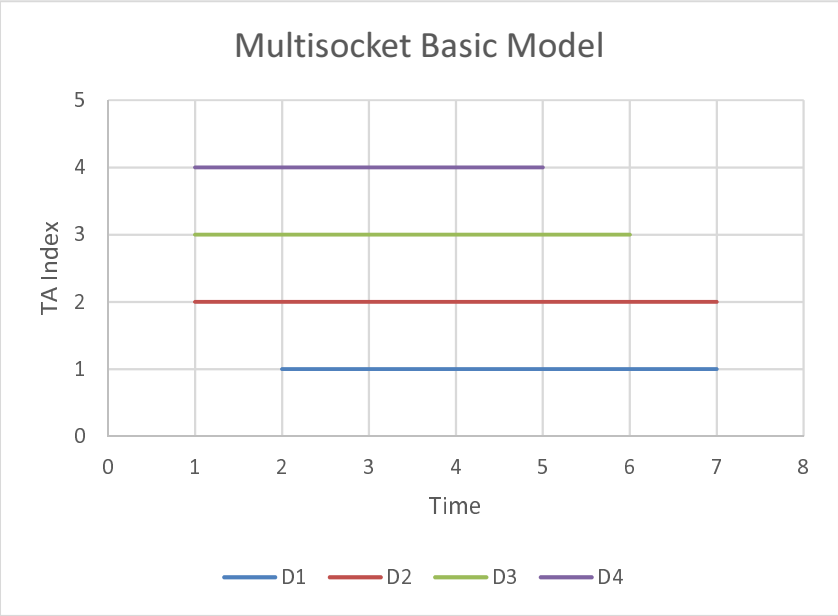

Figure 15.2: Realistic multi-socket model where individual downloads start an end at different points in time

It follows that both the fixed-size and fixed-time approach can be used in analogy to the single-socket case. For instance, an app may use multiple sockets to transfer a given amount of data. Each socket is tasked to transfer a certain fixed percentage of a total data volume, and then closes; this would be a multiple fixed-size approach.

Alternatively, each socket works on a best-effort basis. The app creates and uses a pre-defined or automatically chosen number of sockets until all the content has been transferred.

If the diagnostic goal is to determine the network's performance limit with a given number of sockets, the test would need to end when the first socket has completed its designated transfer.

With reference to Figure 15.2 and the basic rules as given above, the valid time window for a transaction would have different values depending on the purpose of the test:

- If the goal is to test only full multi-socket operation, the time window would start at T=2.
- In case of a user perspective with a given application, the valid time window would start at T=1 if the model contains a set-up phase, and at T=0 if it is full end to end.
- If the test is fixed-time, and the intended time window ends at a given time Tx (before T=5), the valid time window would end at Tx.
- Otherwise (or in hybrid mode), the end of the valid time window would be at T=5 if the purpose is to make a full multi-socket test and this socket terminates intentionally by the app's characteristics (result: successful), or if socket no. 4 has dropped (result: not successful).
- If the purpose is to perform a user-perspective test, and none of the socket drops prematurely, the end of the valid time window would be at T=7.

If the test is about performance with a given number of sockets, and the actual socket count becomes smaller than this number, the test can be ended prematurely to optimize data yield, as waiting for the rest of the sockets to complete would not change the verdict for this transaction anymore.

All other elements of throughput measurement methodology can be used 1:1 - the number of bytes transferred will be the sum of all bytes in all sockets. From an implementation point of view, this means that even if there is an event-based mechanism at the bottom of measurement data delivery - such as creating an event each time a certain amount of data has been received or transmitted - this has to be translated to a data delivery at fixed points in time.

16 Comparability and reproducibility

In analogy to ETSI TS 102 250 [i.2], the information required for a particular set of QoS parameters will be sufficient to reproduce the measurement.

There are many valid ways of doing performance measurements using different platforms and evaluation methods, each with a specific set of parameters. The information elements listed subsequently are just examples, assuming that the actual set of information is a matter of professional judgement and care. In any case, the guiding question for the selection of parameters to be reported is 'what is required to reproduce the measurement?':

- Testing system (Endpoint A); platform-relevant information such as HW platform, operating system.
- Methods and respective parameters for the test.
- Access/interface to the network under test (any elements of the data path between endpoint A and the NUT).
- Set-up and parameters of endpoint B.

NOTE: In case that endpoint B is a live element (e.g. a public server or web site), it cannot be guaranteed or controlled that the configuration is the same. However, reasonable care should be taken to make sure that the configuration is understood as well as possible. This may include, for instance, a discussion of the possible impact and dimension of deviations. Such a discussion should be aimed at putting expected deviations of results into perspective, e.g. compare them to sample count related statistical errors and other influences.

Some examples of parameters to be recorded are:

- Method used (e.g. fixed-size, fixed-time, other).
- File size or data volume limit.
- Type of transfer (UDP, http/ftp, etc.; respective traffic profile characteristics such as packet sizes, etc.).
- Radio access technologies under whose coverages the measurement are performed.
- Timeout/time window for measurements using the fixed-time method.
- Number of parallel transfers.
- Point of Observation (PoO) for data size and temporal characteristics of data capture (e.g. if the information is sampled with a fixed time interval or created event-driven after a certain amount of data has been transferred). This will also allow to understand if measurements taken with this PoO represent usable throughput (sometimes termed 'goodput') or if values include retransmissions. Likewise, this information enables to assess possible impact of granularity due to packet size and time values.
- Evaluation: time windows t1, t2 or algorithm to select them.
- Evaluation: selection method for throughput samples.

To assess statistical errors, information on the number of samples is required. Even if basic considerations indicate that two QoS parameter values can actually be compared, this information is required to determine if numerical values are in plausible vicinity or not.

17 Summary and conclusion

At closer analysis, it becomes clear that the seemingly simple field of performance measurement, in particular for throughput or 'network speed', requires a great deal of professional care. If values obtained by different measurement tools will be compared in a meaningful way, it is not sufficient to look at the measurement method or the equipment used. It is also necessary to consider the methods used for computation of raw data. The present document is an approach to define these requirements and provide a framework for reliable measurements.

Annex A: Bibliography

- ETSI TS 102 250 (all parts): "Speech and multimedia Transmission Quality (STQ); QoS aspects for popular services in mobile networks".
- IETF RFC 8337: "Model-Based Metrics for Bulk Transport Capacity".

History

Document history		
V1.1.1	October 2018	Publication