

# 1

## What is QoS?

### 1.1 QoS Definition

According to ISO 8402, the word quality is defined as “the totality of characteristics of an entity that bear on its ability to satisfy stated and implied needs”. ISO 9000 defines quality as the degree to which a set of inherent characteristics fulfils requirements. ITU-T (Recommendation E.800 [ITU-TE.800]) and ETSI [ETSI-ETR003] basically defines Quality of Service (QoS) as “the collective effect of service performance which determine the degree of satisfaction of a user of the service”. As stated in [ETSI-TR102], IETF considers QoS as the ability to segment traffic or differentiate between traffic types in order for the network to treat certain traffic flows differently from others. QoS encompasses both the service categorization and the overall performance of the network for each category.

Concerning the network viewpoint, QoS is the ability of a network element (e.g. an application, host or router) to have some level of assurance that its traffic and service requirements can be satisfied. QoS manages bandwidth according to application demands and network management settings.

The term “QoS” is used in different meanings, ranging from the users’ perception of the service to a set of connection parameters necessary to achieve particular service quality. The QoS meaning changes, depending on the application field and on the scientific scope. Reference [Gozdecki03], starting from the terminology concerning QoS in IP networks, defines some reference points about the QoS issue. The authors, mentioning [Hardy01], identify three types of QoS: intrinsic, perceived and assessed. Intrinsic QoS is directly provided by the network itself and may be described in terms of objective parameters as, for instance, loss and delay. Perceived QoS (P-QoS) is the quality perceived by the users; it heavily depends on the network performance but it is measured by the “average opinion” of the users. Mean Opinion Score (MOS) methods are often used to perform the measure of the quality: users assign an MOS rating to the application they are evaluating as follows: 1 – bad, 2 – poor, 3 – fair, 4 – good, 5 – excellent. The MOS is the arithmetic mean of all the individual scores, and can range from 1 (worst) to 5 (best) [MOS].

Even if there is a strict connection with the objective metrics provided by the network, the user does not necessarily perceive an objective performance increase (or decrease), in correspondence of an intrinsic QoS variation. References [Adami01] and [Adami02] report a performance evaluation by using MOS values along with measures of objective metrics and show that the correspondence is not always straightforward. The topic is of extreme importance and deserves great attention. Control schemes applied to telecommunications, which will receive a great deal of attention within this book, imply a big economic effort, both in the design and in the implementation phase. If, for example, the application of a specific control algorithm leads to a performance improvement of 20% concerning the performance metric (e.g. a decrease of 20% of the lost packets while traversing a network), it appears as an excellent result and, actually, from an objective viewpoint, it is. But, is it perceived by the real users? Can human eyes and ears appreciate the improvement? Is the economic effort provided to implement the control architecture worthwhile? A practical case study taken from the direct experience of the author within the framework of the activity performed in the 3-year project "Integration of Multimedia Services on Heterogeneous Satellite Networks", funded by the Italian Space Agency (ASI) and carried out by the Italian National Consortium for Telecommunications (CNIT), may help understand. The case study presents an experimental approach to provide a guaranteed QoS over a satellite network based on the Internet Protocol (IP). The aim is to get a proper environment for data, voice and video transmission to be used for tele-working and distance learning through videoconference. Two Local Area Networks (LANs), located in Site A and Site B, are connected through a satellite link at 2 Mbps; the LAN located in Site C, where no satellite station is available, is connected to the network in Site A by using Integrated Services Digital Network (ISDN). The gateways among the LANs and the external parts (satellite stations or ISDN) are represented by routers, which are network devices operating at IP level. The details of the trials are not relevant in this context. The focus is on the relation between objective performance parameters and subjective perception, which is related both to the characteristics of human perception and to the type of application. In short, the videoconference service is experimentally provided through the network both applying two bandwidth reservation control mechanisms within the intermediate routers to protect audio flows and not implementing any bandwidth protection scheme. The three alternatives are identified as Full-Control, Light-Control and No-Control. The implementation complexity is decreasing with the same order. The percentage of lost packets is measured during the tests as well as the MOS of the participants. The bit rate of the video is incremented step by step. Tables 1.1 and 1.2 contain a small portion of the results in [Adami02] but sufficient to have an idea of the problem.

**Table 1.1** Objective metrics

Bit/rate	Video packet loss percentage			Audio packet loss percentage		
	No-control [%]	Light-control [%]	Full-control [%]	No-control [%]	Light-control [%]	Full-control [%]
128	0.05	0.40	0.05	0	0.10	0.09
256	21	7.88	6.50	31.45	0.40	0.23
384	34.73	22	7.74	38.70	0.08	0
512	46.05	33.10	12.92	50.98	0	0

**Table 1.2** Subjective metrics

Bit/rate	Videoconference MOS			Video MOS			Audio MOS		
	No-control	Light-control	Full-control	No-control	Light-control	Full-control	No-control	Light-control	Full-control
128	3.5	3.75	3.75	2.75	3	3.25	4	4.5	4.5
256	1.75	3.75	3.75	1.75	3.25	3.5	1.5	4.5	4.5
384	1.5	3	3.5	1	2	3.25	1	4.5	4.5
512	1	2.75	3.5	1	1.75	3.25	1	4	4.5

The first table reports the values of Objective Metrics: video and audio packet loss during the videoconference sessions. Table 1.2 shows the Subjective Metric MOS corresponding to the configurations of Table 1.1, divided for media: audio, video and overall videoconference service. Even if the number of users of the system is not sufficient to have precise measures, the values can provide interesting indications. Concerning video it is important to note that if the percentage of lost packets is about 10%, the perception of the users is still quite good (well above 3). If, for example, the video bit rate is set to 256 Kbps, a big effort to implement a Full-Control is not necessary and the implementation of Light-Control is sufficient to provide a satisfying service. It is even clearer if the MOS of the overall service is analysed: at 256 Kbps the evaluation of videoconference is exactly the same for both Full-Control and Light-Control. This measure also allows to make a further comment while looking at the audio packet loss and at the corresponding MOS metrics. The Light-Control can guarantee a low audio packet loss (approximately the same as the Full-Control); there is a very positive perception of the audio (4.5), which compensates the slight performance decrease of the video in the user's perception. Actually, audio performance is dominant in videoconference and most of the overall evaluation is due to it.

Additionally, if video loss is above 20%, the video perception is very low, but the overall videoconference evaluation (supported by the audio result) for Light-Control is still acceptable. Only when the video performance is really damaged (e.g. video bit rate of 512 Kbps and measured loss of 33%), the users of the service are annoyed and the videoconference evaluation is below 3. In this case, the use of a complex and expensive control is fully justified. Concerning the measure of audio, it is sufficient to have a look at the performance in presence of No-Control: a packet loss of about 30% is sufficient to make the overall service completely useless.

Which is the message of this limited example? Not necessarily an objective performance improvement is perceived by the users and the perception is not always the same because it heavily depends on the type of service.

The last type of QoS reported in [Hardy01] concerns assessed QoS. It is referred to the will of a user to keep on using a specific service. It is related to P-QoS and also depends on the pricing mechanism, level of assistance of the provider and other marketing and commercial aspects. For example, a performance decrease may be surely tolerated by a user if a service is free, but the same decrease will raise criticism if the user is paying for it.

At the moment, most of the QoS provision is offered in terms of intrinsic (objective parameters) QoS by using a Service Level Specification (SLS) which is "a set of parameters and their values which together define the service offered to a traffic" [RFC3206]. SLS is

a separated technical part of “a negotiated agreement between a customer and the service provider on levels of service characteristics and the associated set of metrics” [Gozdecki03, ITU-T-Y.1241], which is the commonly adopted definition of a Service Level Agreement (SLA). An example of SLS is represented by the Asynchronous Transfer Mode (ATM ) Traffic Contract [ATM-Forum96] that is composed of traffic parameters and descriptors, along with a set of QoS parameters. SLS used in this work includes the type of traffic (e.g. Premium VBR, Mission critical and best effort); traffic description and conformance testing (packet dimension, application peak and average rate, and, if requested, maximum burst and bucket size); and performance guarantees (packet loss rate, packet transfer delay and packet delay jitter). Another possible example of SLS which the author, together with his research group, proposes for military tactical networks is reported in Table 1.3. It includes also the feature of Multi-Level Precedence and Pre-emption (MLPP), which is a peculiar characteristic of military voice switches but, in military environment, is recommendable also for data traffic. It is of topical importance in tactical environment and establishes a level of priority for calls by using 4 bits ([Polk01, Kingston00]). To make an example taken from [Polk01], normal calls are classified as “Routine”; a lower-level command traffic uses “Priority” and “Immediate” levels. Brigade, battalion and division (field grade traffic) is assigned “Immediate” and, in some cases, “Flash”. Corps commanders use “Flash”. Presidents and Joint Chiefs use “Flash override”. It is used to drop a call when a higher priority connection tries entering the network and the bandwidth is not sufficient for both. Actually, military networks are challenging for communication engineers because they include all the services of interest for civil communications but require different performance requirements in terms of loss, delay and jitter. There are also specific applications that concern the military world. In short, SLS specifies the service from the technical viewpoint. The Service Level Agreement (SLA) includes also non-technical issues [Sarangan2006] as well as pricing and device/network capabilities. For example, customers may be available to relax technical constraints in dependence of the price applied; on the other hand, service providers may offer the same service at different prices depending on the current network utilization. Concerning device and network capabilities, the SLA may also indicate the access technology that the customer can use, for example Ethernet, Wireless LANs, GSM, UMTS, Ad-hoc Network, GPRS and Satellite Access. Even if the SLS does not change, SLA should also consider these aspects that have a relevant impact on the customer choices.

**Table 1.3** Possible example of SLS

Service Level Specification	Range
Connection type	Constant Bit Rate (CBR)/Variable Bit Rate (VBR)/Best-effort
Scope	End-to-end
Connection identification	Identifier or sequence of identifiers
Traffic description and conformance testing	Peak rate/bucket size for peak rate/maximum burst size
Performance guarantees	Packet loss rate/packet transfer delay/packet delay jitter
Multi Level Precedence and Pre-emption (MLPP)	Not applied/routine/priority/immediate/flash/flash override

**Table 1.4** Operative example of SLS

Service Level Specification	Range
Connection type	Constant Bit Rate (CBR)
Scope	End-to-end
Connection identification	Sequence of identifiers
Traffic description and conformance testing	Peak rate = 64 Kbps/bucket size for peak rate = 512 bytes/maximum burst size not applicable
Performance guarantees	Packet loss rate = 1%/packet transfer delay = 250 ms/packet delay jitter = 30 ms
Multi Level Precedence and Preemption (MLPP)	Priority

An operative example applying the SLS defined above is contained in Table 1.4, where a “lower-level command voice call” for military communication is reported.

## 1.2 Applications

Which applications need QoS? The answer is simple: all the applications that require a specific level of assurance from the network. The answer does not give any idea about the amount of applications that need QoS. Some of them are listed below: basic services for information transfer for both backbone and access networks, assured database access to retrieve information, telemedicine (transmission of clinical tests, X-rays, electrocardiograms and magnetic resonance), tele-control (remote control of robots in hazardous environments, remote sensors and systems for tele-manipulation), bank and financial operations, purchase and delivery, tele-learning, telephony, videoconferences and applications for emergencies and security.

Having very different characteristics, each mentioned application deserves a specific degree of service, defined at the application layer. Several standardization bodies have tried to define service categories (also called “QoS classes”, to be intended at application layer).

ITU-T (in Recommendation Y-1541 [ITU-T-Y.1541]) suggests a definition of QoS classes (for the IP world) that is summarized in Table 1.5.

The ETSI Project TIPHON [ETSI-TR102] proposes an alternative QoS class definition, reported in Table 1.6.

Concerning Broadband-Integrated Services Digital Network (B-ISDN), ITU-T defines a set of service categories, which is reported in Table 1.7 [Onvural94].

**Table 1.5** ITU-T Y-1541 QoS classes

QoS class	Characteristics
0	Real-time, jitter sensitive, highly interactive
1	Real-time, jitter sensitive, interactive
2	Transaction data, highly interactive
3	Transaction data, interactive
4	Low loss only (short transactions, bulk data, video streaming)
5	Traditional applications of default IP networks

**Table 1.6** TIPHON QoS classes (from [ETSI-TR102])

QoS class	Components	QoS characteristics
Real-time conversational (telephony, teleconference, videophony and videoconference)	Speech, audio, video, multimedia	Delay and delay variation sensitive, limited tolerance to loss and errors, constant and variable bit rate
Real-time streaming (e.g. audio and video broadcast, surveillance, graphics)	Audio, video, multimedia	Tolerant to delay, delay variation sensitive, limited tolerance to loss and errors, variable bit rate
Near real-time interactive (e.g. web browsing)	Data	Delay sensitive, tolerant to delay variation, error sensitive, variable bit rate
Non-real-time background (e.g. e-mail and file transfer)	Data	Not delay and delay variation sensitive, error sensitive, best effort

**Table 1.7** B-ISDN services

Categories	Applications (examples)
Conversational	Videoconference, video surveillance, high speed document communication (file transfer, fax, images, sound)
Messaging	E-mail and video e-mail, chat
Retrieval	High speed document retrieval (video, still images, sounds, transaction data)
Distribution <i>without</i> user-individual presentation control	Document and video distribution, Pay TV, radio
Distribution <i>with</i> user-individual presentation control	Full-channel broadcast TV and videography

Derived from the general categories reported in Table 1.7 and consequently to the standardization of ATM as the technology for implementing B-ISDN, the ATM Forum defines five ATM service categories, reported in Table 1.8 [McDysan99].

Each application mentioned at the beginning of this sub-section may be inserted in the classifications reported above but, besides the general definition reported, each application

**Table 1.8** ATM forum service categories

ATM service categories	Representative applications	QoS characteristics
CBR (Constant Bit Rate)	Circuit emulation	Low cell delay variation, low loss
rt-VBR (real time Variable Bit Rate)	Video on demand	Moderate cell delay variation, low loss
nrt-VBR (non-real-time Variable Bit Rate)	Packet traffic	Moderate loss
ABR (Available Bit Rate)	Adaptable rate sources	Low loss
UBR (Unspecified Bit Rate)	Best-effort traffic	No requirements

(or, in this context, more exactly, each user) needs a detailed specification in terms of traffic descriptors and intrinsic QoS parameters to allow a proper service provision by the network.

### 1.3 QoS Metrics

A further step is to associate objective QoS requirements to QoS traffic classes generically defined above. Concerning the IP environment, the QoS objective metrics mostly used [ITU-T-Y.1540] are as follows:

- IPLR – IP Packet Loss Ratio
- IPTD – IP Packet Transfer Delay
- IPDV – IP Packet Delay Variation (known as Jitter)
- IPER – IP Packet Error Ratio.

Another metric often considered is the skew, which is the average value of the difference of the delays measured by packets belonging to different media, as, for example, voice and video within a videoconference service. In this case, if the skew is large, there is no synchronization between voice and video with the general effect of a bad dubbing.

IP Packet Loss Ratio is the performance measure used in the examples in Tables 1.2 and 1.3. Even if, as said, the problem needs to be solved for each specific application, it is also important to give a range of QoS requirements for traffic classes. Possible end-to-end performance-metric upper bounds are reported in Table 1.9, associated with QoS classes in [ITU-T-Y.1541]. This association is introduced in [ITU-T-Y.1541]. The nature of the objective performance parameter is defined in [ITU-T-Y.1541] as follows: IPTD – upper bound on the mean IPTD; IPDV – upper bound on the  $1-10^{-3}$  quintile of IPTD minus the minimum IPTD; IPLR – upper bound on the packet loss probability; IPER – upper bound. “U” stands for “Unspecified” (actually meaning “Unbounded”, in this context).

The values reported in Table 1.9 apply to IP public networks and are upper bounds to mean quantities. Actually, each single company and service provider may also offer different end-to-end performance. In other words, taking class 0 for example, it means that, for a generic real-time, jitter sensitive, and highly interactive application, the following performance-metric limits should be guaranteed: mean end-to-end transfer delay below 0.1s, mean jitter below 0.05s, mean packet low rate below  $10^{-3}$  and mean packet error rate below

**Table 1.9** IP QoS classes and objective performance-metric upper limits

QoS Class	Characteristics	IPTD	IPDV	IPLR	IPER
0	Real time, jitter sensitive, highly interactive	100 ms	50 ms	$1 \times 10^{-3}$	$1 \times 10^{-4}$
1	Real time, jitter sensitive, interactive	400 ms	50 ms	$1 \times 10^{-3}$	$1 \times 10^{-4}$
2	Transaction data, highly interactive	100 ms	U	$1 \times 10^{-3}$	$1 \times 10^{-4}$
3	Transaction data, interactive	400 ms	U	$1 \times 10^{-3}$	$1 \times 10^{-4}$
4	Low loss only (short transactions, bulk data, video streaming)	1 s	U	$1 \times 10^{-3}$	$1 \times 10^{-4}$
5	Traditional applications of default IP networks	U	U	U	U

**Table 1.10** Examples of multimedia applications' QoS requirements

Application	IPTD	IPDV	IPLR
Data acquisition from sensors	100 ms	50 ms	0
Radar traces	20 ms	1–3 ms	$10^{-3}$
Weapon control	20 ms	10 ms	0
Sensor control	20 ms	10 ms	0
Voice	250 s	30 ms	$10^{-2}$
Video streaming	5–10 s	U	$2 \cdot 10^{-2}$

$10^{-4}$ . The recommendation does not refer to specific applications but defines upper limits for traffic classes.

Clearly, within the scenario defined above, it is topical to have specific upper-bound indications for each application. Table 1.10 tries to give them for a group of applications concerning IPTD, IPDV and IPLR. The IPER threshold defined in Table 1.9 is still applicable. The IPTD for voice may be extended to 400 ms. Interaction is still possible in conformity with Table 1.9 indications.

## 1.4 The Concept of Traffic Flow and Traffic Class

There is no unique definition of traffic flow in the literature. If it is not differently indicated in the text, a flow is considered here as a packet stream associated with a precise user service. So, if traffic is differentiated “per-flow”, it means that each single user is identified. To better clarify this aspect, the term “user flow” is often used in the book. A traffic class is considered a group of flows associated with a common identifier. A traffic class should contain traffic with similar characteristics and performance requirements. “Per-class” traffic separation is simpler but coarser. It is also called “per-aggregate” granularity, in the text. The identification feature will be the object of deep discussion in Chapters 3 and 4 and a distinction factor for the QoS architectures presented in Chapter 6.