Quality of Service

QoS, when measured in communications networks, indicates the overall performance of the connectivity. For measuring the network QoS, the most common QoS statistics include delay/latency, jitter (delay variation), available data rate, and packet loss. As statistics already indicate, QoS is always a measure carried out between two points of interest in the network. Qosium measures QoS from network traffic level, indicating the QoS experienced by applications.

Table of Contents

“Regarding communications, the only thing that matters to the end-user, whether a human or a machine, is the connection quality.”

1. What is QoS and Why Is It Important

A communications system is at its best when the user doesn’t notice its existence while using networked applications and services. Thus, an ideal data communications system would have an unlimited data rate and a constant undelayed and lossless delivery of data packets. In that case, you can think QoS to be indefinitely good. But, as we all know, this is a utopia, so we must be satisfied with something much less. That is not an issue, however, since the meaning of QoS is always application specific. It is enough that the QoS of a communications system is high enough for the applications and services to run smoothly. The requirements vary a lot: a communications path whose quality satisfies one application fully, can be a disaster to another.

For example, a VoIP conversation has very small data rate requirements, and also some packet losses are tolerated except in high compression ratio codecs. The delay starts to get annoying only after some 100 – 200 ms. In online gaming, in contrast, that delay level would already be far too much for games requiring fast reactivity. Then again, industrial automation applications typically require both very low delay and zero packet loss to work safely and efficiently. On the other hand, over-the-top streaming media services such as Netflix and YouTube demand throughput capacity rather than low delays and packet loss levels. While delay and packet loss will also slow down streaming services and bulk data transmission, content buffering in the end device takes mostly care of the sporadic QoS variation in the connectivity.

Is maximum throughput performance QoS? For bulk data transfer, it mostly is, but generally not: it represents only one feature of QoS. The tools that allow you to measure your connection’s speed only tell how much capacity there’s in reserve in your network connection at that time towards the test server. It may give you a hint that a bulk data transfer or a high-quality video stream runs smoothly. But it tells you little about how applications requiring low delay, low jitter, and low packet losses would work. A Ping test often accompanies maximum data speed measurements. It measures *RTT*, a representation of two-way joint delay, but it is only for that Ping application. Some other applications having completely different packet length profiles, packet rate variations, and protocols over the same connection likely experience different delays. Consider if your connection gives sporadic one-way delay spikes for one packet per thousand on average, a high-rate video stream contains these spikes every second. If we assume that Ping’s data packets experience the same kind of spiky delay behavior, being not always the case, with standard settings, it takes hundreds of seconds, on average, to notice the spikes by Ping, and thousands to make a more deterministic perception.

In telecommunications terminology, you often see QoS mentioned in the context of data prioritization, where it refers to the QoS target rather than the realized QoS. For example, many wireless systems are equipped with QoS class definitions to manage different traffic flows differently. There are dedicated bearer definitions for different types of traffic and subscriptions in mobile networks. DiffServ is one commonly known standard for classifying and managing IP traffic. While these solutions attempt to provide the application traffic with satisfactory QoS, they cannot guarantee that. The classification and the following data prioritization can favor defined traffic types over the others, but they cannot battle, e.g., low radio link quality efficiently. In addition, the methods generally do not work end-to-end but just for a limited part of the total network path. Thus, even though these methods help, you still don’t know how well it went, i.e., what kind of QoS did your applications get over the network path.

For end-users, often, the only thing that matters is that the application used works as it should. The application works when it gets the QoS it requires from the network. The application does not care about the underlying network technology. It can be wired or wireless, as long as delays, delay variation, and packet loss stay within tolerable boundaries and the realized data transfer rate is enough for the used application. Real-time QoS measurement enables you to monitor how well the network can serve your applications and how usable your applications are.

2. Typical QoS Statistics

The table below introduces the main QoS statistics. Application QoS demands determine which of them play the essential roles. As the statistics indicate, they are such that you always need to have a reference point over which they can be calculated, i.e., QoS is always determined for the path between two network points. For instance, calculating the delay for an IP packet necessitates that you know when that packet has been sent on the other end of the measurement path. As this is not simple to be carried out, the typical way has been to measure RTT. It, however, cannot tell you what has the delay behavior been in the sent and receive directions, which can sometimes vary a lot. If the packet involved in the round-trip measurement is lost, you cannot know which direction caused this loss. Having one-way statistics helps you analyze network problems in more detail. In addition, as discussed earlier, RTT is typically an *active measurement*, where the traffic to be measured is particularly generated for the purpose. In order to measure real applications' QoS, *passive QoS measurement methods* are required.

Statistic	Description
Delay	Time difference between the time data is transmitted to the moment it is received by the other measurement point. Sometimes delay is also called *latency*.
Jitter	Time how much delay differs between sequential packets
Packet loss ratio	Ratio between successfully transmitted and lost packets on the measurement path
Connection break	Time duration from the moment a packet loss is detected to a moment packet is successfully transferred
Available data rate	Indicates how fast the connection can transfer data

3. Glossary

Round-Trip Time

In packet data communications, RTT is the time it takes a packet to be sent from one network point to another and back.

In practical measurement solutions, like Ping, RTT includes the one-way delays of the directional network paths plus the processing delay of the solution. The processing delay is typically considered small or insignificant compared to the delay caused by the network paths. Sometimes RTT is called Round-Trip Delay, RTD.