Quality of Service

In computer networks, the term **Quality of Service**, usually abbreviated as **QoS**, refers to the parameters used to describe the quality of the service provided by a network—such as speed, packet loss and delay—or to the tools and techniques used to achieve a desired level of service quality.

The following sections describe the main attributes used to evaluate the quality of a network connection.

Bandwidth

In computing, **bandwidth** refers to the amount of data that can be transferred from one point to another within a given period of time.

$$Bandwidth = \frac{Information\ size}{Transfer\ time}$$

Bandwidth is measured in **bits per second**, abbreviated as **bps**. The **bit rate** is the number of bits transferred in one second.

Its most common multiples are:

- **Kilobits per second (Kbps):** one thousand bits per second.
- **Megabits per second (Mbps):** one million bits per second.
- **Gigabits per second (Gbps):** one billion bits per second.

Bandwidth can be used to describe the speed of a host's Internet connection because it indicates how quickly the host can download or upload data.

This speed may vary depending on the type of Internet connection—such as DSL or optical fibre—and several other factors.

In networking, the term bandwidth is often used to indicate the **maximum theoretical data-transfer capacity** of a connection, measured in bits per second.

In this context, saying that a connection provides 100 Mbps means that, under ideal conditions, it can transfer up to 100 million bits per second.

Bandwidth may be either symmetrical or asymmetrical:

- **Symmetrical bandwidth** means that the download speed—also called downstream speed, from the server to the client—and the upload speed—also called upstream speed, from the client to the server—are the same.
- **Asymmetrical bandwidth** means that the download and upload speeds are different.

Many residential ISPs provide asymmetrical connections with higher download speeds than upload speeds. This is because users generally download more content from the Internet than they upload.

Example
ADSL stands for **Asymmetric Digital Subscriber Line**.

This type of connection is asymmetrical because its downstream bandwidth is greater than its upstream bandwidth.

Another important concept is **throughput**.

Throughput is the actual amount of data transferred during a particular period of time.

Unlike theoretical bandwidth, throughput takes real network conditions into account, including:

- Congestion
- Interference
- Distance
- Signal quality
- Packet loss
- Traffic generated by other users

Throughput is therefore normally lower than the available bandwidth, or equal to it only under ideal conditions.

Example
Suppose that we need to download a 20 MB file and the download takes 20 minutes.

Since one byte is equal to eight bits:

- File size: `20 MB × 8 = 160 Mb`
- Transfer time: `20 minutes = 1,200 seconds`

The average throughput is therefore:

`160 Mb ÷ 1,200 s ≈ 0.133 Mbps`

This is approximately **133 Kbps**.

The theoretical minimum transfer time for content of a particular size can be calculated from the bandwidth using the following formula:

$$Transfer\ time = \frac{Information\ size}{Bandwidth}$$

Example
Suppose that we want to transfer a 10 MB file using a connection with a bandwidth of 5 Mbit/s.

First, we convert megabytes into megabits:

```
10 MB × 8 = 80 Mbit
```

The theoretical transfer time is therefore:

```
80 Mbit ÷ 5 Mbit/s = 16 seconds
```

Latency

Latency is the time between sending a packet, request or signal and its reception—or the corresponding response—at the destination.

It represents the delay between the beginning of an action and its actual execution or response.

Latency is measured in **milliseconds**, abbreviated as **ms**, and is affected by several factors within a network.

The main factors include:

- **Connection transmission rate:** A higher transmission rate can reduce the time required to place data onto a link, although bandwidth is only one component of overall latency.
- **Physical distance:** Latency increases with the physical distance between the source and destination. Long-distance connections, such as connections between continents, may therefore have higher latency.
- **Network congestion:** When a network carries too much traffic, packets may remain in queues before being forwarded towards their destination.
- **Infrastructure quality:** The quality and reliability of routers, cables and other network components can affect overall latency.
- **Network protocols:** Some network protocols require more processing or communication exchanges than others and may therefore influence latency.

It is important to understand that **bandwidth and latency are not the same thing**.

A high-bandwidth connection can transfer large amounts of data every second, but this does not necessarily mean that it has low latency.

Latency also depends on physical distance, the route followed by packets, congestion and the processing time required by network devices.

Latency can significantly affect activities that require rapid responses, such as online gaming and video calls.

For example, high latency in an online game may create a noticeable delay between the player’s action and the game’s response, negatively affecting the gaming experience.

Low latency is generally preferable, particularly for real-time applications, because it provides smoother and more responsive communication.

Network operators and engineers attempt to reduce latency by optimising infrastructure, implementing advanced technologies and using efficient protocols.

Example
A connection with a latency of approximately **15–30 ms** may feel highly responsive in online games.

A latency of **80–100 ms** may still be usable, but the delay can become more noticeable in real-time applications.

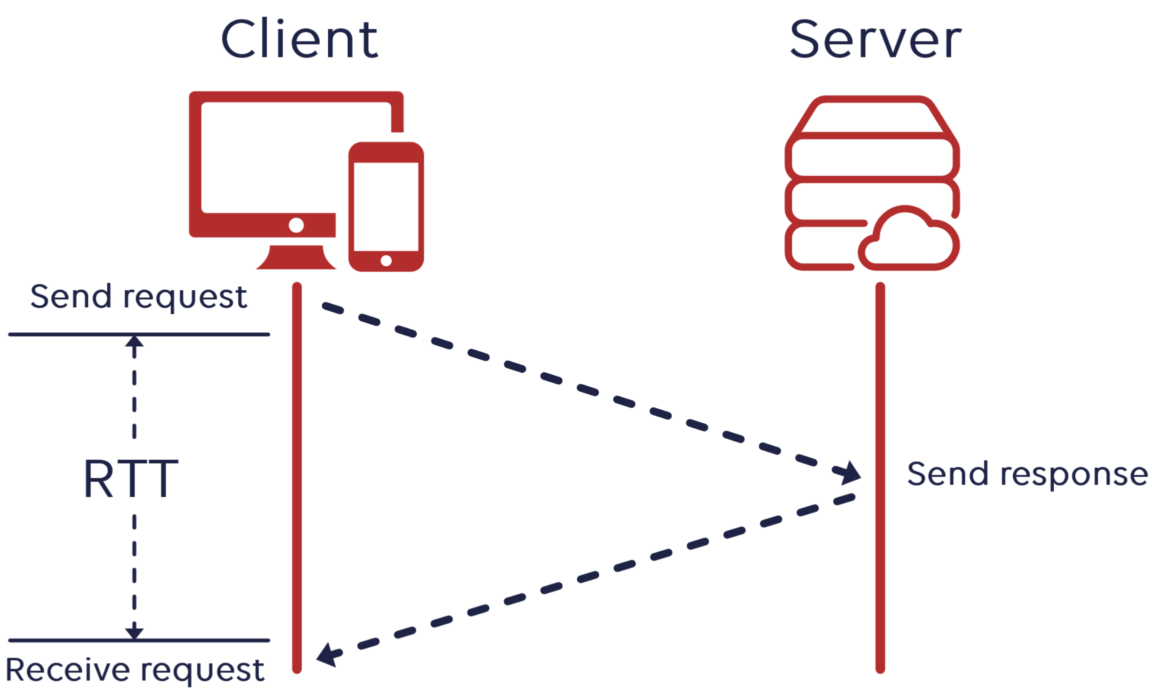
The **Round-Trip Time**, abbreviated as **RTT**, is a measurement related to latency.

It represents the total time required for a data packet to travel from a source to a destination and for the response to return to the source.

In other words, RTT is the total time needed to complete a round trip between two points in a network connection.

In an ideal symmetrical network, RTT may be approximately twice the one-way latency because it includes both the journey to the destination and the return journey.

In real networks, however, this relationship is not always exact because the outgoing and return paths may experience different delays.



Jitter

Jitter is another important networking concept, particularly for real-time applications such as Internet voice calls and videoconferencing.

Jitter represents the variation in packet latency as packets travel from a source to a destination.

- If packets arrive with very different delays, jitter is high.
- If they arrive with similar and regular delays, jitter is low.

Jitter can create problems for real-time applications by causing irregularities in the transmission of voice or video during a call.

When jitter is low, the packet flow is more regular and the user experience is better because there are fewer interruptions and synchronisation problems.

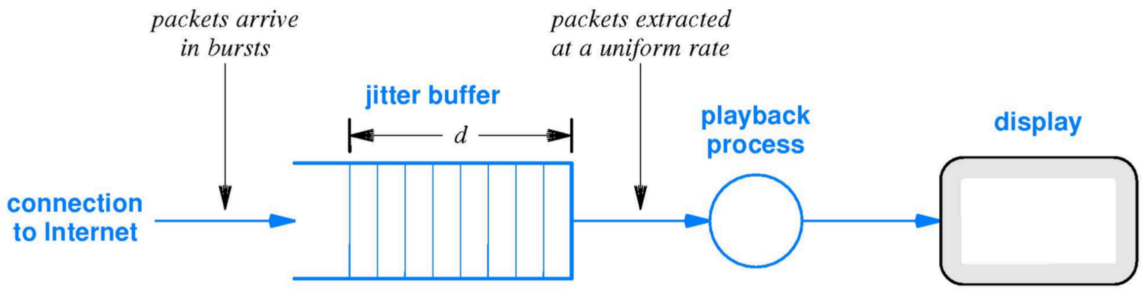
Possible causes of jitter include:

- **Network congestion:** When a network is overloaded, packets may be delayed or arrive out of sequence.
- **Variable network paths:** Packets following different routes may experience different delays.
- **Queue management:** The way network devices manage packet queues can affect jitter. If packets accumulate in queues, their delays may vary considerably.

To reduce jitter and improve the quality of real-time communication, network providers and application developers use techniques such as:

- Traffic prioritisation
- Buffers
- More stable network routes

These techniques attempt to reduce latency variation and maintain a regular data flow.



One way of quantifying jitter is to calculate the **standard deviation of packet delay**.

Standard deviation is a statistical measurement describing the dispersion or variability of a set of values around their average.

In other words, it indicates how much the individual values tend to differ from the mean value.

Example
If the average value in a dataset is 5 and the standard deviation is 0, every value in the dataset is equal to 5.

Buffer

A **buffer** is an area of temporary memory used to compensate for differences in speed between two devices or processes exchanging data.

In practice, a buffer stores data while it waits to be processed or transmitted. It acts as a cushion between the component producing the data and the component consuming it.

Imagine water flowing from a tap faster than it can be used: an intermediate container is needed to store it temporarily.

In a digital system, this container is the buffer: a temporary area of memory that compensates for speed differences and timing irregularities.

Example
Buffers are used in many different situations:

- **Printing:** A file is temporarily buffered before being sent to the printer.
- **Networks:** A router stores incoming packets in buffers before forwarding them.
- **Digital audio:** Audio recorded in real time is buffered so that it can be processed without interruptions.
- **Writing to a disk:** Data is placed in a buffer before being written to the storage device.

An important use of buffers on the Internet is limiting the effects of temporary variations in data reception.

In audio or video streaming, for example, a buffer loads part of the content in advance to prevent interruptions during playback.

When a user starts a video on a service such as YouTube or Netflix—or begins listening through a service such as Spotify—the system does not necessarily play every piece of data as soon as it arrives.

Instead, it stores part of the content in a **local buffer**, meaning a temporary area of memory.

Once the buffer contains enough data, playback begins. Meanwhile, new data continues to arrive and refill the buffer.

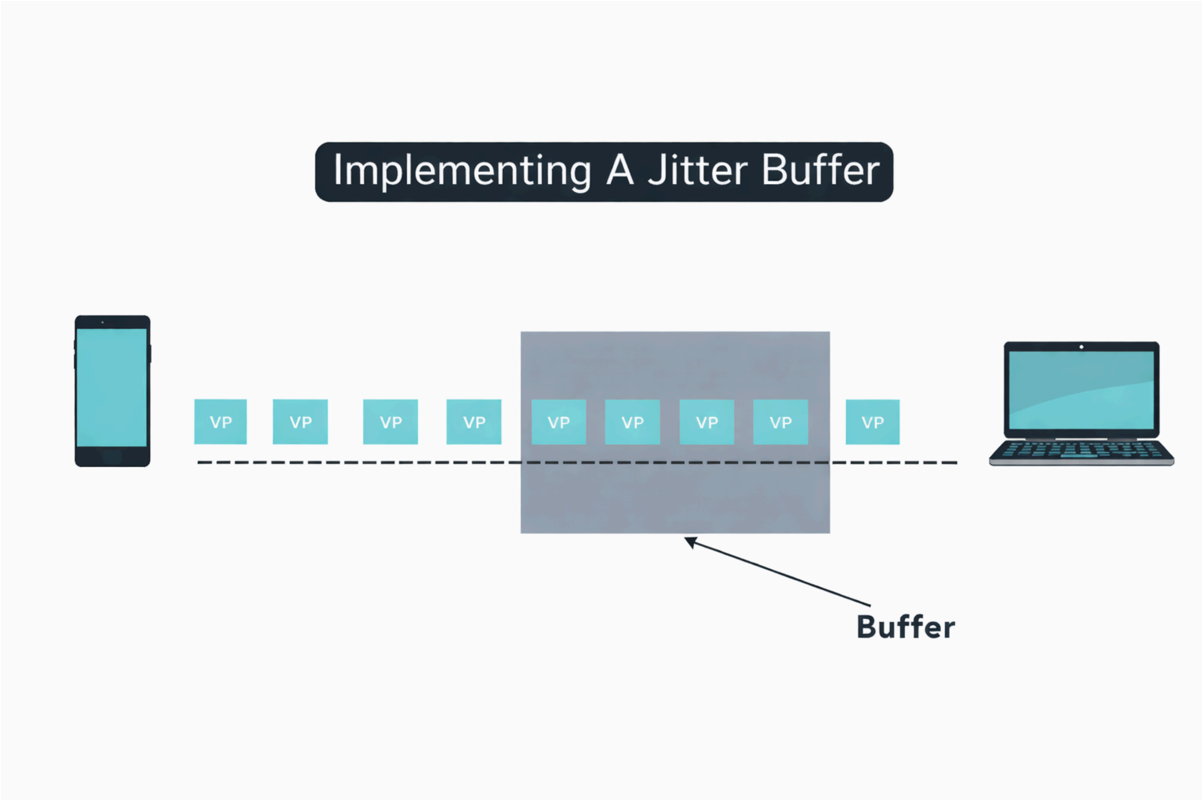
This process protects playback against temporary network slowdowns and allows content to play more smoothly even when the data does not arrive at a perfectly regular rate.

In real-time streams, such as voice calls and video calls, a more specific type of buffer called a **jitter buffer** is used.

A jitter buffer is designed to compensate for variations in packet arrival times while keeping latency as low as possible.

The main difference is:

- In traditional streaming, the system can store more data in advance, even if this introduces a longer delay.
- In a video call, the delay must remain minimal because participants need to interact in real time.



The diagram shows how a buffer stores packets and then sends them to the destination device as a flow with more consistent timing.

A buffer is therefore used to compensate for jitter and make playback more regular.

- High jitter may require a larger buffer.
- However, an excessively large buffer introduces additional latency because the system must wait longer for it to fill.

A balance is therefore required. This is often managed through dynamic jitter-buffer algorithms.

Quality of Service for Media

Adequate Quality of Service is important when delivering media because poor network performance can degrade the media itself.

QoS requirements vary according to the type of media:

- **Non-continuous media:** With non-continuous media such as text and images, latency may affect how long the user waits before the content appears. Once the data has been received, however, it can be displayed without following a precise timing sequence.
- **Continuous media:** With continuous media such as audio and video, the data must not only arrive but also arrive regularly. An initial delay may be acceptable for

non-interactive content because it can be compensated for with a buffer. Highly variable packet delays are more problematic because audio may be interrupted and video may stutter.
- **Live streaming:** The content is produced and transmitted almost in real time. A buffer may still be used, but it cannot be excessively large, or the viewer will see the event with a significant delay.
- **Interactive media:** In interactive applications such as online games and video calls, latency must be even lower because users expect an immediate response to their actions.

The difficulty of managing media QoS can be divided into four levels of increasing complexity:

1. Managing non-continuous media such as images and text.
2. Streaming previously recorded content, or **stored media**, such as watching a video on YouTube.
3. Streaming content produced in real time, or **live media**, such as Internet radio. The main difficulty in designing streaming and teleconferencing applications is network delay. Audio and video must be presented at regular times to remain useful.
4. Managing interactive applications such as video games and teleconferencing. Excessive delays prevent communication from remaining genuinely interactive.

Ping

A simple way to analyse certain aspects of Quality of Service is to use **ping**.

Ping is a network command used to check whether a host can be reached through an IP network and to measure its response time.

It can also help identify packet loss by showing whether some requests receive no response.

The name comes from the sound produced by sonar systems in submarines. Just as a sonar echo checks whether something responds, ping checks whether a host responds through the Internet or a LAN.

When ping is executed:

- The computer sends a special packet to an IP address or domain name.
- If the destination device is reachable and configured to respond, it sends a reply packet.

Ping can be run through the operating system's terminal.

For example:

```
ping google.com
```

A Windows system may produce a response similar to:

```
Risposta da 142.250.184.14: byte=32 tempo=24ms TTL=118
```

Where:

- **bytes = 32:** The size of the data included in the ping packet.
- **time = 24 ms:** The time required to receive the response, meaning the measured RTT.
- **TTL = 118:** The remaining **Time to Live** value in the received IP packet. Each router normally reduces this value by one, preventing packets from travelling indefinitely through the network.