Qualità del servizio

Nel campo delle reti, il termine qualità del servizio o più semplicemente ***QoS*** (dall'inglese Quality of Service) è utilizzato per indicare i parametri usati per caratterizzare la qualità del servizio offerto dalla rete (ad esempio la velocità, perdita di pacchetti, ritardo, ecc..), o gli strumenti o tecniche per ottenere una qualità di servizio desiderata.
Sotto un elenco degli attributi principali per definire la qualità di una connessione di rete.

Banda

La banda, in informatica, si riferisce alla quantità di dati che possono essere trasferiti da un punto all'altro in un dato periodo di tempo.

$$Banda = \frac{Dimensione\ delle\ informazioni}{Tempo\ di\ trasferimento}$$

La banda viene misurata in bit al secondo, indicati con bps. **Il bit rate** è la quantità di bit trasferiti in un secondo. I suoi multipli più comuni sono kilobit al secondo (*Kbps*, mille bits al secondo), megabit al secondo (*Mbps*, un milione di bits al secondo) e gigabit al secondo (*Gbps*, un miliardo di bits al secondo).
La larghezza di banda può riferirsi alla velocità di Internet di un host perché indica quanto velocemente possono scaricare o caricare dati su Internet.
Questa velocità può variare a seconda del tipo di connessione Internet (ad esempio, *DSL*, fibra ottica, ecc…) e di altre variabili.

Nel linguaggio delle reti, il termine banda viene spesso usato per indicare la capacità massima teorica di trasferimento dati di una connessione, misurata in bit al secondo.
In questo contesto, parlare di una connessione da 100 Mbps significa indicare che, in condizioni ideali, quella connessione può trasferire fino a 100 milioni di bit al secondo.

La larghezza di banda può essere simmetrica o asimmetrica:

- Una larghezza di banda simmetrica significa che le velocità di download (downstream, la velocità di trasferimento dei dati dal server verso il client) e upload (upstream, la velocità di trasferimento dei dati dal client verso il server) sono le stesse.
- Mentre una larghezza di banda asimmetrica ha velocità di download e upload diverse.

Molti *ISP* domestici offrono connessioni asimmetriche, con velocità di download più elevate rispetto alle velocità di upload, questo perché statisticamente siamo portati più a scaricare contenuti da internet piuttosto che caricare nuovi contenuti.

Esempio
L'*ADSL* sta per Asymmetric Digital Subscriber Line. Tale tecnologia di connessione è asimmetrico perché la banda del downstream è maggiore della banda del upstream.

Altro concetto importante è Il **throughput**.
Il throughput è la quantità effettiva di dati che viene trasferita in un certo intervallo di tempo. A differenza della banda teorica, il throughput tiene conto delle condizioni reali della rete, come congestione, interferenze, distanza, qualità del segnale, perdita di pacchetti e traffico generato da altri utenti.
Quindi il throughput è sempre minore della banda o al limite uguale.

Esempio
Devo scaricare un file di 20 MB (un byte equivale a 8 bit) e ci impiego 20 minuti.
$Throughput = \frac{20\ MB}{20\ m} = \frac{20\ 000\ 000\ B}{20 * 60\ s} = \frac{20\ 000\ 000 * 8\ b}{20 * 60\ s} = \frac{160\ 000\ 000\ b}{1200\ s} = 133\ 333\ bps \approx 133\ kbps$

Dalla larghezza di banda si può calcolare il tempo di trasferimento minimo teorico di un contenuto di una certa dimensione in byte.
consideriamo la formula inversa della banda:

$$tempo\ di\ trasferimento = \frac{Dimensione\ delle\ informazioni}{Banda}$$

Esempio
Supponiamo di voler trasferire un file di 10MB (10 megabyte) su un dispositivo con banda di 5 Mbit/s.
Il tempo di trasferimento T sarà dato da:
$Tempo\ di\ trasferimento = \frac{10\ MB}{5\ Mbps} = \frac{10\ 000\ 000\ B}{5\ 000\ 000\ bps} =$
$= \frac{10\ 000\ 000 * 8\ b}{5\ 000\ 000\ bps} = \frac{80\ 000\ 000\ b}{5\ 000\ 000\ bps} = 16s$

Un utile strumento per misurare il throughput medio sui dispositivi è in questo link.

Latenza

La **latenza** (o tempo di latenza; in inglese: latency) è il tempo che intercorre tra l'invio di un pacchetto, di una richiesta o di un segnale e la sua ricezione o risposta da parte della destinazione: rappresenta il ritardo che si verifica tra l'inizio di un'azione e la sua effettiva esecuzione o risposta.

La latenza è misurata in millisecondi (*ms*) ed è influenzata da diversi fattori all'interno di una rete. Alcuni dei principali fattori che contribuiscono alla latenza includono:

- Velocità della connessione: Una connessione Internet più veloce tende a ridurre la latenza poiché i pacchetti di dati vengono trasmessi più rapidamente.
- Distanza fisica: La latenza aumenta con la distanza fisica tra la sorgente e la destinazione dei dati. Ad esempio, le connessioni a lunga distanza, come quelle tra continenti, possono avere latenze più elevate.
- Congestione di rete: Quando la rete è sovraccarica di traffico, i pacchetti possono subire ritardi nell'essere instradati verso la destinazione desiderata.
- Qualità dell'infrastruttura: La qualità e l'affidabilità dei router, dei cavi e di altri componenti di rete possono influenzare la latenza complessiva.

- Protocolli di rete: Alcuni protocolli di rete sono più efficienti di altri e possono ridurre la latenza.

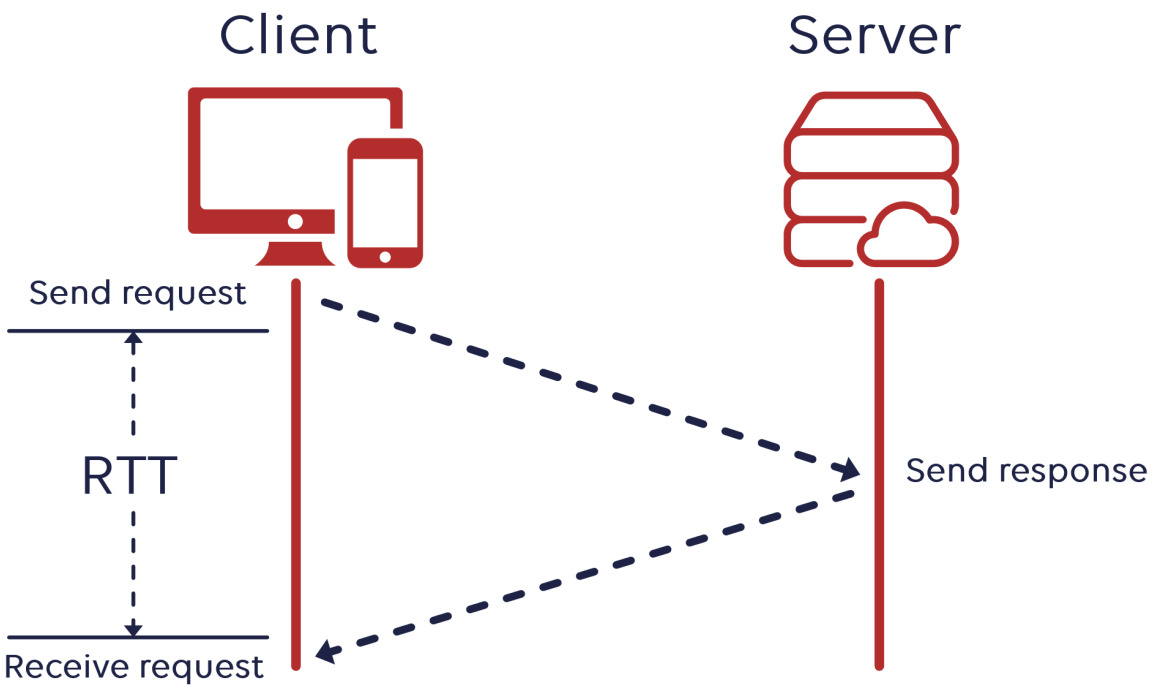
E' importante sottolineare che banda e latenza non sono la stessa cosa.
Una connessione con molta banda può trasferire grandi quantità di dati al secondo, ma non è detto che abbia una latenza bassa.
La latenza dipende anche dalla distanza fisica, dal percorso seguito dai pacchetti, dalla congestione e dai tempi di elaborazione nei dispositivi di rete.

La latenza può avere un impatto significativo su diverse attività, specialmente in applicazioni che richiedono tempi di risposta rapidi, come i giochi online o le videochiamate.
Ad esempio, una latenza elevata nei giochi online potrebbe causare ritardi tra l'azione dell'utente e la risposta del gioco, influenzando negativamente l'esperienza di gioco.

In generale, una latenza bassa è preferibile, soprattutto per le applicazioni in tempo reale, poiché consente un funzionamento più fluido e responsivo della rete.
Gli operatori di rete e gli ingegneri di rete cercano costantemente di migliorare la latenza attraverso l'ottimizzazione dell'infrastruttura, l'implementazione di tecnologie avanzate e l'uso di protocolli efficienti.

Esempio
Una connessione con latenza intorno a 15–30 ms può risultare molto reattiva per giochi online. Invece una latenza di 80–100 ms è ancora utilizzabile, ma può rendere più percepibile il ritardo nelle applicazioni in tempo reale.

Il **Round Trip Time** (*RTT*, tempo di andata e ritorno) è una misura correlata alla latenza in e rappresenta il tempo impiegato da un pacchetto di dati per andare da un punto di origine a un punto di destinazione e tornare indietro al punto di origine.
In altre parole, il *RTT* è il tempo totale necessario per completare un giro (andata e ritorno) tra due punti di una connessione di rete.
In una rete ideale e simmetrica, il *RTT* può essere considerato circa il doppio della latenza solo andata, perché comprende sia il tempo necessario per raggiungere la destinazione sia il tempo necessario per ricevere la risposta. Nelle reti reali, però, questa relazione non è sempre precisa, perché il percorso di andata e quello di ritorno possono avere ritardi diversi.



Un utile strumento per misurare l'*RTT* tra il vostro dispositivo e una pagina web è in questo link.

Jitter

Il **jitter** è un altro importante concetto particolarmente rilevante per applicazioni in tempo reale, come le chiamate vocali su Internet o le videoconferenze.
Il jitter rappresenta la variazione nella latenza dei pacchetti di dati mentre viaggiano da una sorgente a una destinazione sulla rete.

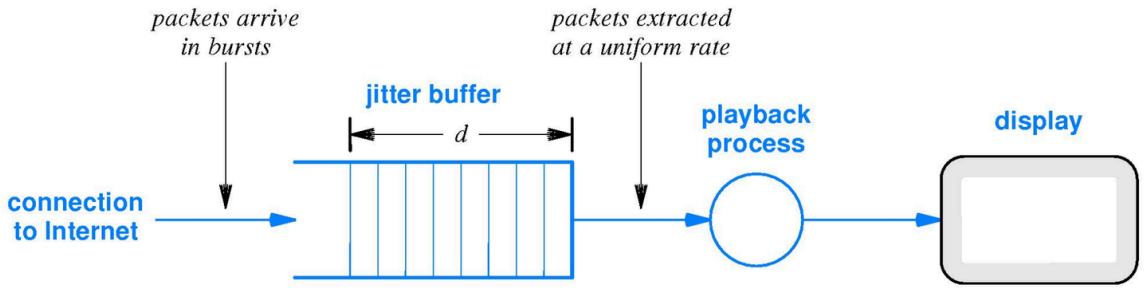
- Se i pacchetti arrivano con ritardi molto diversi tra loro, il jitter è alto.
- Se invece arrivano con ritardi simili e regolari, il jitter è basso.

Il jitter può essere problematico per le applicazioni in tempo reale perché può causare fenomeni come ritardi nella trasmissione della voce o delle immagini durante una videochiamata.
Se il jitter è basso, il flusso dei pacchetti è più uniforme e l'esperienza dell'utente è migliore, in quanto ci sono meno interruzioni o problemi di sincronizzazione tra i partecipanti.

Le cause del jitter possono essere diverse, tra cui:

- Congestione di rete: Quando una rete è sovraccarica di traffico, i pacchetti possono essere ritardati o ricevuti fuori sequenza, causando jitter.
- Percorsi di rete variabili: Se i pacchetti devono attraversare percorsi diversi sulla rete, possono subire latenze diverse, contribuendo al jitter.
- Gestione della coda: Nei dispositivi di rete, la gestione della coda dei pacchetti può influenzare il jitter. Se i pacchetti vengono accumulati in una coda troppo grande o troppo piccola, il loro ritardo può variare.

Per affrontare il jitter e migliorare la qualità delle comunicazioni in tempo reale, i fornitori di servizi di rete e gli sviluppatori di applicazioni implementano diverse tecniche, come la prioritizzazione del traffico, l'utilizzo di buffer (vedremo fra poco questa tecnica) o la scelta di percorsi di rete più stabili, al fine di ridurre la variabilità nella latenza e mantenere un flusso regolare dei dati.



Il jitter possiamo vederlo come la deviazione standard della latenza.
La deviazione standard è una misura statistica che rappresenta la dispersione o la variabilità di un insieme di dati rispetto alla loro media. In altre parole, indica quanto i valori di un dato campione tendono a differire dalla media campionaria.

Esempio
Se la media di un insieme di dati è 5 e la deviazione standard è 0 vuol dire che tutti i dati sono equivalenti a 5.

Un utile strumento per misurare il jitter tra il vostro dispositivo e una pagina web è in questo link.

Buffer

Il termine buffer indica una porzione di memoria temporanea usata per compensare differenze di velocità tra due dispositivi o processi che si scambiano dati.
In pratica, il buffer accumula dati in attesa di essere elaborati o trasmessi, agendo da "cuscinetto" (da cui il nome) tra chi produce i dati e chi li consuma.

Immagina un rubinetto che esce più veloce di quanto il bicchiere possa assorbire → serve un contenitore intermedio.
Nei sistemi digitali, quel contenitore è il **buffer**: una RAM temporanea che aiuta a compensare differenze di velocità e irregolarità temporali.

Esempio
Alcuni esempi concreti di uso del buffer:

- Stampa: Il file da stampare viene bufferizzato prima di essere inviato alla stampante.
- Reti: Un router usa buffer per immagazzinare pacchetti in arrivo prima di inoltrarli.

- Audio digitale: I suoni registrati in tempo reale vengono bufferizzati per essere elaborati senza interruzioni.
- Scrittura su disco: I dati da salvare vengono messi in un buffer e poi scritti su disco.

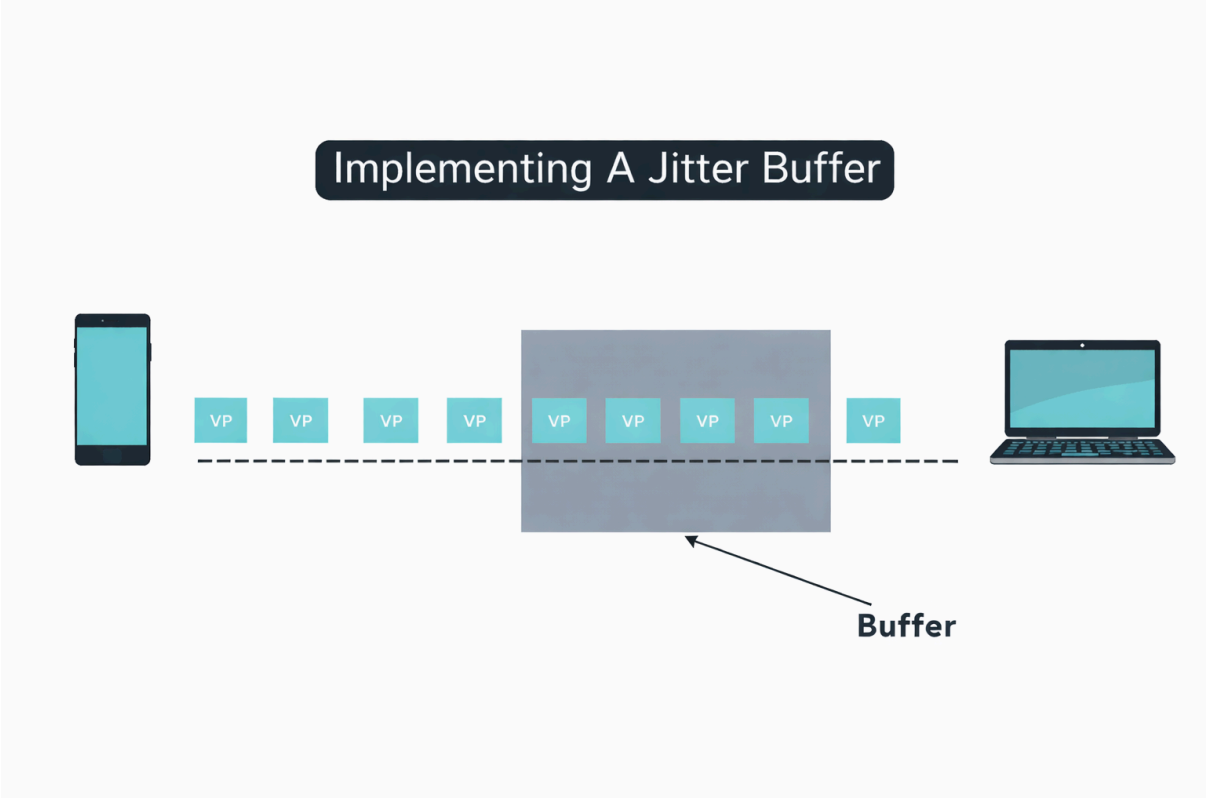
Un uso importante del buffer su Internet è limitare gli effetti delle variazioni temporanee nella ricezione dei dati. Per esempio, nello streaming video o audio, il buffer serve a caricare in anticipo una parte del contenuto, in modo da evitare interruzioni durante la riproduzione.

Quando l'utente avvia un video, per esempio su YouTube, Netflix o Spotify, il sistema non riproduce immediatamente ogni dato appena ricevuto.
Invece, accumula una porzione del contenuto in un buffer locale, cioè in una memoria temporanea. Quando il buffer contiene abbastanza dati, inizia la riproduzione.
Nel frattempo, nuovi dati continuano ad arrivare e a riempire il buffer.

Questo processo protegge da rallentamenti temporanei della rete e permette una riproduzione più fluida, anche se i dati non arrivano sempre con regolarità.

Nei flussi in tempo reale, come chiamate vocali e videochiamate, si parla più specificamente di **jitter buffer**.
Un jitter buffer è un buffer progettato per compensare le variazioni nei tempi di arrivo dei pacchetti, mantenendo però la latenza più bassa possibile.

La differenza principale è:
- Nello streaming tradizionale il sistema può permettersi di accumulare più dati in anticipo, introducendo anche un ritardo maggiore.
- In una videochiamata, invece, il ritardo deve rimanere minimo, perché gli utenti devono poter interagire in tempo reale.



In questo schema si vede come funziona il buffer: conserva i pacchetti in modo poi di mandare un flusso con latenza costante al dispositivo finale

Il buffer è usato per compensare jitter e rendere regolare la riproduzione.

- Se il jitter è alto, serve un buffer più grande.
- Ma un buffer troppo grande introduce latenza (si deve aspettare tanto prima che il buffer si riempia), quindi serve un compromesso intelligente, spesso gestito da algoritmi jitter buffer dinamici.

Qualità di servizio nei media

Avere una qualità di servizio adeguata per visualizzare i media è importante visto che una qualità di servizio bassa comporta un degrado del media stesso.

La qualità di servizio può avere specifiche diverse in base alle diverse tipologie di media:

- *Media non continui:* Nei media non continui, come testo e immagini, la latenza può influenzare il tempo di attesa prima della visualizzazione. Tuttavia, una volta ricevuti i dati, il contenuto può essere mostrato senza dover rispettare un ritmo temporale preciso.
- *I media continui*: Nei media continui, come audio e video, non conta solo che i dati arrivino, ma anche che arrivino con regolarità. Un ritardo iniziale può essere accettabile se il contenuto non è interattivo, perché può essere compensato con un buffer. Diventa invece più problematico se i pacchetti arrivano con ritardi molto variabili, perché l'audio può risultare interrotto e il video può andare a scatti.
- Live streaming: Il contenuto viene prodotto e trasmesso quasi in tempo reale. Anche qui si può usare un buffer, ma non può essere troppo grande, altrimenti l'utente vedrebbe l'evento con un ritardo eccessivo.
- *Media interattivi*: Nelle applicazioni interattive, come videogiochi online e videochiamate, il ritardo deve essere ancora più basso, perché l'utente si aspetta una risposta immediata alle proprie azioni.

.

Per le difficoltà sulla qualità di servizio nei media ci sono quattro livelli di difficoltà sotto visti in ordine crescente:

- La gestione di media non continui, come immagini e testo.
- La gestione di streaming di contenuti registrati (*stored media*), come vedere un video su youtube.
- Lo streaming dei contenuti prodotti in tempo reale (*live media*), come la radio. Il problema principale nel progettare applicazioni di streaming e di teleconferenza è il ritardo di rete. Audio e video richiedono presentazione in tempo reale, il che significa che per essere utili devono essere inviati sulla rete a tempi fissati.
- Applicazioni interattive, come i videogiochi o le teleconferenze. In questo caso grandi ritardi implicano che le chiamate che dovrebbero essere interattive non lo sono più.

Ping

Un modo semplice per analizzare la qualità del servizio è il **ping**.
Il **ping** è un comando di rete usato per verificare se un host è raggiungibile tramite una rete *IP* e per misurare il tempo di risposta. Può anche aiutare a individuare eventuali perdite di pacchetti, osservando se alcune richieste non ricevono risposta.
Il nome deriva dal suono del sonar nei sottomarini: come un'eco sonora verifica se qualcosa risponde, il ping verifica se un host risponde su Internet o in una *LAN*.

- Quando si esegue un ping, il computer invia un pacchetto speciale a un indirizzo *IP* o a un nome di dominio.
- Il dispositivo di destinazione, se raggiungibile, risponde con un altro pacchetto.

Il ping lo si può scrivere come comando del terminale del sistema operativo. Per esempio se si scrive:

```
ping google.com
```

Il risultato sarà:

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

Dove:

- byte = 32 → dimensione del pacchetto inviato.
- tempo = 24 *ms* → tempo impiegato per ricevere la risposta, cioè il suo *RTT*.
- *TTL* = 118 → *Time To Live*, quanti "salti" di router può fare il pacchetto prima di scadere