



Telecommunications

▼ 1.0 - Introduzione

▼ 1.1 - Introduzione alle telecomunicazioni

Le **telecomunicazioni** (comunicazioni a distanza) si basano sull'invio e la ricezione di segnali contenenti l'informazione desiderata. Nell'ambito delle comunicazioni elettriche questi segnali corrispondono a tensioni e correnti variabili nel tempo, spesso ottenute convertendo in segnali elettrici segnali di altra natura, ad esempio sonori o visivi, tramite degli appositi trasduttori (microfoni, telecamere). Una volta inviati e ricevuti, questi segnali vengono di nuovo convertiti nella forma originaria da altri trasduttori (altoparlanti, schermi televisivi). Un generico segnale elettrico viene descritto matematicamente mediante funzioni reali o complesse del tempo $x(t)$.

Più in generale i segnali si possono classificare secondo molti criteri diversi:

- **Segnali tempo continui o tempo discreti**

I primi sono definiti su tutto l'asse dei tempi, come nel caso dei segnali analogici, i secondi solo per un'infinità numerabile di istanti t_n con n intero, posti ad un intervallo T . Il segnale in questo caso non è altro che una successione di valori $x_n = x(t_n)$, come nel caso dei messaggi digitali.

- **Segnali continui o discreti nei valori**

Nei primi, fissato un generico istante t_1 , la funzione $x = x(t_1)$ può assumere con continuità tutti i valori di un determinato intervallo, $[-M, M]$, come nei segnali analogici, nei secondi solo una quantità numerabile, di norma finita, come nei segnali digitali (due soli, nel caso binario).

- **Segnali deterministicici e segnali aleatori**

Nel primo caso è noto l'andamento prima della trasmissione, cioè a priori; ad esempio perché esso segue una funzione nota analiticamente come una sinusoide, un'onda quadra, ecc. Appartengono di norma a questa categoria i segnali di prova generati in laboratorio. Nel secondo caso l'andamento è noto solo dopo la trasmissione, cioè a posteriori. Appartengono a questa categoria sia il rumore elettrico (ad esempio il fruscio che si può facilmente percepire in assenza di segnale utile, nei dispositivi analogici, come la radio), sia, necessariamente, tutti i segnali

non noti al ricevitore. Si noti che se il ricevitore conoscesse a priori il segnale che sta per ricevere non ci sarebbe trasmissione di informazione dalla sorgente per definizione ("se so già che cosa stai per dirmi, non mi darai nessuna informazione"). Nonostante in questo corso ci occuperemo prevalentemente dei segnali deterministicici, i segnali più importanti che fanno parte della teoria dell'informazione corrispondono a quelli aleatori.

- **Segnali ad energia finita o a potenza finita**

I segnali possono anche essere classificati secondo un criterio "energetico". Nel caso dei segnali ad energia finita converge (ad un valore diverso da zero, altrimenti il segnale è evidentemente nullo) l'integrale che definisce l'energia di un segnale, nel secondo l'energia è infinita ma converge invece quello che ne definisce la potenza media. Si noti che esistono anche segnali che non appartengono a nessuna delle due categorie.

▼ 1.2 - Richiami sui numeri complessi

I **numeri complessi** sono un'estensione dei numeri reali, nata per trovare tutte le soluzioni delle equazioni polinomiali. In particolare la seguente equazione $x^2 = -1$ non ha soluzioni reali, perché i quadrati dei numeri reali non possono essere negativi. Si definisce perciò l'unità immaginaria "i" che gode della proprietà $i^2 = -1$.

Un generico numero complesso "z" viene quindi visto come somma di una parte reale "a" e di una parte immaginaria "b", dove a e b sono numeri reali e i è l'unità immaginaria:

$$z = a + ib$$

Come i numeri reali sono in corrispondenza biunivoca con i punti di una retta, quelli complessi sono in corrispondenza con i punti di un piano: al numero complesso $z = a + ib$ si associa il punto di coordinate cartesiane (a, b) .

Somma e prodotto di numeri complessi

A rigore un numero complesso è definito come una coppia ordinata di numeri reali (a, b) . Sull'insieme di queste coppie vengono quindi definite le operazioni di somma e prodotto come:

$$(a, b) + (c, d) = (a + c, b + d)$$
$$(a, b)(c, d) = (ac - bd, ad + bc)$$

Con queste due operazioni, l'insieme dei numeri complessi \mathbb{C} risulta essere un campo.

- $(a, 0) = a$
- $(0, 1) = i$

Rappresentazione con coordinate cartesiane e coordinate polari

Così come un punto in un piano può essere definito in termini di coordinate cartesiane o di coordinate polari, così è possibile anche per i numeri complessi, che possono essere definiti anche in termini di modulo r (distanza del punto dall'origine degli assi) ed argomento φ (angolo di rotazione rispetto all'asse reale). Notazione cartesiana e notazione polare sono equivalenti in quanto è possibile passare facilmente dall'una all'altra. In particolare si ha che

$$\begin{aligned} a &= r \cos \varphi \\ b &= r \sin \varphi \end{aligned}$$

e, inversamente

$$r = \sqrt{a^2 + b^2}$$

$$\varphi = \begin{cases} \operatorname{arctg} \frac{b}{a} & a > 0 \\ \operatorname{arctg} \frac{b}{a} + \pi & a < 0 \end{cases}$$

Ricordando infine la fondamentale formula di Eulero $e^{i\varphi} = \cos \varphi + i \sin \varphi$, possiamo quindi mettere in evidenza l'equivalenza fra rappresentazione cartesiana e polare di un numero complesso:

$$z = a + ib = re^{i\varphi}$$

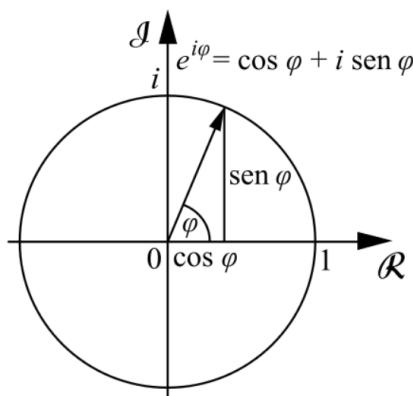


Fig.1 Piano complesso e formula di Eulero.

La formula di Eulero consente anche di esprimere le funzioni coseno e seno in termini di somma e differenza di funzioni esponenziali:

$$\cos \varphi = \frac{e^{i\varphi} + e^{-i\varphi}}{2}$$

$$\sin \varphi = \frac{e^{i\varphi} - e^{-i\varphi}}{i2}$$

o come parte reale ed immaginaria di un unico esponenziale:

$$\cos \varphi = Re\{e^{i\varphi}\}$$

$$\sin \varphi = Im\{e^{i\varphi}\}$$

Sinusoidi e fasori

In tutta la prima parte del corso avremo a che fare con segnali sinusoidali (o scomponibili in somma di segnali sinusoidali), per cui è conveniente richiamare subito il legame che essi hanno con i fasori e con i numeri complessi. Si consideri una funzione sinusoidale, assumendo il tempo come variabile indipendente.

$$x(t) = A \cos(\omega t + \vartheta)$$

La formula di Eulero e le sue derivate permettono di esprimere la nostra funzione sinusoidale sia come somma di due funzioni esponenziali complesse:

$$x(t) = A \cos(\omega t + \vartheta) = \frac{A}{2} e^{j(\omega t + \vartheta)} + \frac{A}{2} e^{-j(\omega t + \vartheta)}$$

sia come parte reale di una qualsiasi delle due precedenti, moltiplicata per due (ad esempio la prima),

$$x(t) = A \cos(\omega t + \vartheta) = Re\{A e^{j(\omega t + \vartheta)}\}$$

Il termine fasore indica una funzione del tipo $A e^{j(\omega t + \vartheta)}$, mentre la costante complessa $A e^{j\vartheta}$ viene chiamata numero complesso rappresentativo del fasore.

▼ 2.0 - Analisi di Fourier di segnali tempo-continui

L'analisi di Fourier consiste nello scomporre un segnale generico in una somma (in generale infinita) di sinusoidi, allo scopo di estendere a segnali generici la soluzione dell'equazione della trasmissione del calore applicabile a segnali sinusoidali, già nota. In termini generali, essa consiste in una fase di analisi, in

cui si cerca di scomporre un segnale generico in termini sinusoidali, ed in una fase di sintesi, in cui il segnale viene rappresentato come somma di termini sinusoidali.

Nell'ambito di questo corso la vedremo applicata a segnali $x(t)$ definiti nel "dominio del tempo" (la variabile indipendente t rappresenta il tempo). In questo caso essa fornisce una rappresentazione alternativa, detta nel "dominio delle frequenze", dei medesimi segnali; per questo motivo viene anche detta "analisi in frequenza". In altre parole, per ampie categorie di segnali sarà possibile dare due rappresentazioni equivalenti, una nel tempo ed una in frequenza, esattamente come per i numeri complessi è possibile affiancare la rappresentazione polare a quella cartesiana. Il passaggio dalla rappresentazione nel tempo a quella in frequenza viene anche detto trasformazione di Fourier (in senso lato), il passaggio inverso antitrasformazione di Fourier (sempre in senso lato).

La trasformazione di Fourier in senso lato assume nomi diversi, ed impiega formule leggermente diverse, a seconda dei segnali a cui si applica. La tabella sotto vuole costituire un'anticipazione di quanto verrà trattato in seguito a questo proposito, per sottolinearne il legame unitario. Allo studente che non ne coglie il legame, esse possono facilmente sembrare una bable di formule diverse, difficili da ricordare. La difficoltà in realtà è nella comprensione del significato dei due domini, tempo e frequenza, che richiede tempo per maturare, non nella memorizzazione delle formule.

Tabella 1: Strumenti dell'Analisi di Fourier

Nome	Segnale nel tempo	Rappresentazione nelle frequenze
Sviluppo in serie di Fourier (tre forme)	Segnali periodici; $x(t)=x(t+T)$.	$\{c_n\}; \{A_n\}, \{\varphi_n\}; \{a_n\}, \{b_n\}$.
Trasformata di Fourier (ed integrale di Fourier)	Segnali aperiodici; $x(t)$.	$X(\omega); V(\omega), \varphi(\omega)$.
Trasformata di F. di una serie temporale	Serie temporali; $\{x_n\}_{n=-\infty}^{+\infty}$.	$X_s(\omega)$.
Trasformata discreta di Fourier, o DFT (FFT)	Ennupla di valori; $\{x_n\}_{n=0}^{N-1}$.	$\{X_p\}_{p=0}^{N-1}$.

Prima di procedere è necessario ribadire che lo scopo dell'analisi di Fourier è la rappresentazione mediante funzioni sinusoidali (o ad essa riconducibili, come i fasori) di segnali che sinusoidali non sono. Se il segnale di partenza è già sinusoidale, o è un fasore, o è già espresso come somma di sinusoidi e di fasori, è evidentemente un grave errore applicare ad esso l'analisi di Fourier.

▼ 2.1 - Sviluppo in serie di Fourier

Lo sviluppo in serie di Fourier si applica alle funzioni tempo continue $x(t)$, continue o discrete nei valori, periodiche con periodo T .

Forma esponenziale (segnali complessi)

Vi sono tre forme, ed inizieremo da quella esponenziale che è la più generale, essendo applicabili sia alla funzioni reali che a quelle complesse. Sotto alcune condizioni, che vedremo fra poco, la funzione $x(t)$ può essere rappresentata come somma di infiniti fasori, aventi pulsazioni multiple della pulsazione fondamentale $\omega_0 = \frac{2\pi}{T}$ secondo la formula (di sintesi) seguente, detta serie di Fourier (in forma esponenziale):

$$x(t) = \sum_{n=-\infty}^{\infty} c_n e^{jn\omega_0 t}$$

I numeri complessi rappresentativi dei fasori, cioè i "coefficienti" della serie di Fourier in forma esponenziale, sono dati dalla formula (di analisi):

$$c_n = \frac{1}{T} \int_{-\frac{T}{2}}^{+\frac{T}{2}} x(t) e^{-j n \omega_0 t} dt$$

Le formule sopra possono anche essere date in alternativa utilizzando le frequenze, anziché le pulsazioni, semplicemente definendo la frequenza fondamentale come $f_0 = \frac{1}{T}$ e sostituendo $\omega_0 = 2\pi f_0$. Il significato è ovviamente il medesimo.

Condizioni di convergenza di Dirichelet

Se per il segnale periodico $x(t)$ sono soddisfatte le seguenti condizioni (solo sufficienti):

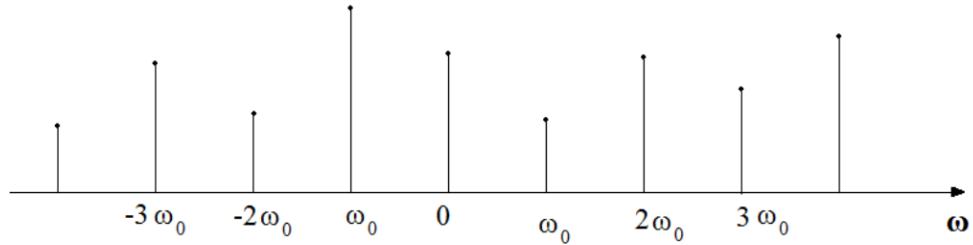
- $x(t)$ è assolutamente integrabile, cioè $\int_0^T |x(t)| dt < \infty$
- Il numero di massimi e di minimi di $x(t)$ (o meglio della parte reale e della parte immaginaria se $x(t)$ è complesso) in un periodo è finito
- Il segnale $x(t)$ è continuo o al più presenta un numero finito di discontinuità di prima specie (cioè il limite sinistro e quello destro sono ovviamente diversi nel punto di discontinuità, ma entrambi finiti)

Allora la serie di Fourier converge in modo puntuale a $x(t)$ dove $x(t)$ è continuo, a $\frac{x(t^+) + x(t^-)}{2}$ nei punti di discontinuità di prima specie.

Spettri di ampiezza e fase bilateri

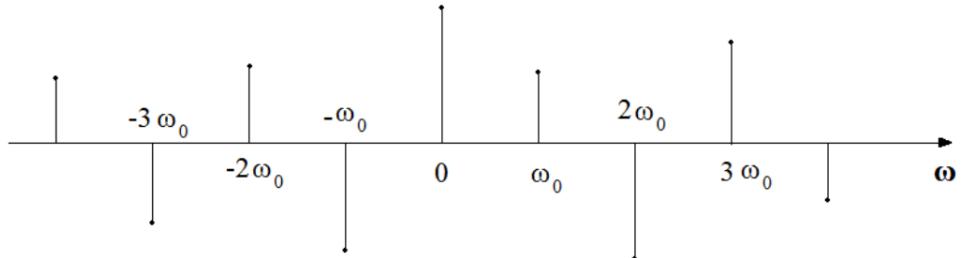
Poiché i coefficienti del segnale sono in generale complessi, non è possibile in generale rappresentarli su un grafico in funzione della pulsazione (o frequenza) associata. Per questo motivo se ne grafica di solito il **modulo**, detto spettro di ampiezza, e l'**argomento**, detto spettro di fase:

$$|\mathbf{C}_n|$$



Spettro di ampiezza bilatero (a righe).

$$\arg\{\mathbf{C}_n\}$$



Spettro di fase bilatero (a righe).

Rappresentazioni monolatere (segnali reali)

Per un segnale reale lo spettro di ampiezza risulta simmetrico rispetto all'origine e
quello di fase antisimmetrico (simmetria hermitiana), cioè:

$$\begin{aligned} |c_n| &= |c_{-n}| \\ \arg\{c_n\} &= -\arg\{-c_n\} \end{aligned}$$

Sfruttando la simmetria hermitiana è possibile manipolare la funzione di sintesi in modo da utilizzare solo le frequenze positive, ottenendo la forma in soli coseni dello sviluppo in serie di Fourier:

$$x(t) = A_0 + \sum_{n=1}^{\infty} A_n \cos(n\omega_0 t - \varphi_n)$$

dove:

$$A_n = \begin{cases} c_0 & n = 0 \\ 2|c_n| & n > 0 \end{cases}$$

$$\varphi_n = -\arg\{c_n\}$$

▼ Dimostrazione

Spezziamo la sommatoria esponenziale di Fourier in tre termini:

$$\begin{aligned} x(t) &= \sum_{n=-\infty}^{-1} c_n e^{jn\omega_o t} + c_o + \sum_{n=1}^{\infty} c_n e^{jn\omega_o t} \\ &= \sum_{n=1}^{\infty} c_{-n} e^{-jn\omega_o t} + c_o + \sum_{n=1}^{\infty} c_n e^{jn\omega_o t} \\ &\stackrel{\text{simmetria Hermitiana}}{=} c_o + \sum_{n=1}^{\infty} 2\operatorname{Re}\{c_n e^{jn\omega_o t}\} = c_o + \sum_{n=1}^{\infty} \operatorname{Re}\{2c_n e^{jn\omega_o t}\} \end{aligned}$$

Operando le seguenti sostituzioni:

$$\begin{aligned} c_o &= A_o \\ 2c_n &= A_n e^{-j\varphi_n} \end{aligned}$$

Otteniamo:

$$x(t) = A_0 + \sum_{n=1}^{\infty} A_n \cos(n\omega_0 t - \varphi_n)$$

Q.E.D.

Analogamente si può pervenire alla terza forma, in seni e coseni (ma senza fasi) definendo i coefficienti monolateri a_n e b_n , anch'essi reali (positivi e negativi), nel seguente modo:

$$\begin{aligned} a_n &= \begin{cases} 2c_0 & n = 0 \\ \operatorname{Re}\{2c_n\} & n > 0 \end{cases} \\ b_n &= -\operatorname{Im}\{2c_n\} \quad n > 0 \end{aligned}$$

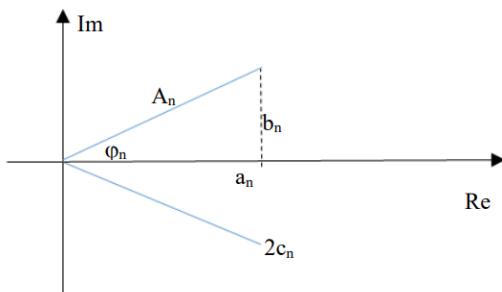
ed operando le sostituzioni:

$$c_0 = \frac{a_0}{2}$$

$$2c_n = a_n - ib_n \quad n > 0$$

ottenendo quindi:

$$x(t) = \frac{1}{2}a_0 + \sum_{n=1}^{\infty} a_n \cos(n\omega_0 t) + \sum_{n=1}^{\infty} b_n \sin(n\omega_0 t)$$



Relazione fra i coefficienti delle tre forme dello sviluppo in serie di Fourier.

I coefficienti a_n e b_n possono anche essere ricavati direttamente (senza passare dalla conoscenza dei c_n) dalle formule seguenti, ottenibili immediatamente dalla formula di analisi e dalle definizioni dei coefficienti stessi:

$$a_n = \frac{2}{T} \int_{-\frac{T}{2}}^{+\frac{T}{2}} x(t) \cos(n\omega_0 t) dt \quad n \geq 0$$

$$b_n = \frac{2}{T} \int_{-\frac{T}{2}}^{+\frac{T}{2}} x(t) \sin(n\omega_0 t) dt \quad n > 0$$

Si noti che se il segnale è pari, la funzione integranda nell'integrale che definisce b_n è dispari (prodotto di una pari e di una dispari), per cui l'integrale si annulla ($b_n=0$) e i coefficienti c_n sono reali; inoltre, nel primo integrale la funzione integranda è pari (prodotto di pari) per cui l'integrale può essere limitato ad un semiperiodo, cosa che a volte può tornare utile negli esercizi:

$$a_n = \frac{4}{T} \int_0^{+\frac{T}{2}} x(t) \cos(n\omega_0 t) dt \quad n \geq 0$$

Analogamente, se $x(t)$ è dispari, $a_n=0$, quindi i c_n sono puramente immaginari, ed infine

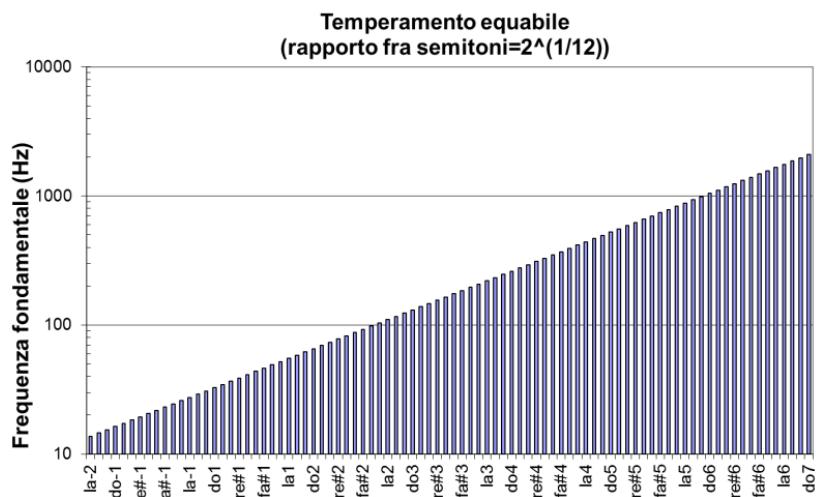
$$b_n = \frac{4}{T} \int_0^{\frac{T}{2}} x(t) \sin(n\omega_0 t) dt \quad n > 0$$

Relazione con la musica

Ci limitiamo pertanto a ricordare che il nostro orecchio funziona come un trasduttore acustico/elettrico, in cui gioca un ruolo fondamentale la "coclea", che di fatto restituisce un diverso stimolo elettrico al variare della frequenza del suono stesso, stimolo che poi viene trasmesso dai nervi e quindi elaborato dal cervello. Iniziamo dal considerare l'emissione di una nota. Il "volume" con la quale una nota è emessa non è altro che la sensazione umana (soggettiva) legata all'intensità (oggettiva) del suono. Il pianoforte si chiama così proprio perché in grado di emettere una nota con intensità diverse (piano o forte), al contrario di altri strumenti a corda come i clavicembali, dove la corda è pizzicata. L'intensità di un suono, trasformato in segnale elettrico, è legata all'ampiezza della variazione del segnale. Di fatto è un fattore moltiplicativo della forma d'onda. Nel dominio delle frequenze esso si traduce in una moltiplicazione di ugual valore di tutti i coefficienti A_n dello sviluppo in serie di Fourier.

L' "altezza" della nota è legata alla frequenza del segnale emesso, ovvero alla frequenza della prima armonica, o armonica fondamentale. Ad ogni nota corrisponde quindi (almeno in un certo tipo di accordatura) una ben specifica frequenza della fondamentale. Dato che la nota emessa non è sinusoidale, ma periodica, avrà non una riga, ma un pettine di righe, poste a frequenza multipla della fondamentale. Ad esempio il "la" del diapason, corrispondente al la della quarta ottava del pianoforte, è associato alla frequenza di 440 Hz. Note superiori avranno fondamentali più alte e quindi pettini più diradati, e viceversa. Il nostro orecchio è in grado di percepire la "distanza" in frequenza fra le righe del pettine, e quindi di rilevare con buona precisione la frequenza del suono emesso. Che differenza c'è quindi fra il la della quarta ottava e quello della quinta? Ad ogni ottava (intervallo di 8 note, da cui il nome) la frequenza di emissione raddoppia (andando verso l'alto), o si dimezza (verso il basso). Il la della quinta ottava avrà quindi frequenza 880 Hz e quello della terza 220 Hz. Il nostro orecchio le riconosce come "la" in entrambi i casi, perché una parte delle righe si trovano nello stesso posto. In particolare andando verso l'alto avrà una riga no ed una riga sì rispetto all'ottava precedente. Capito che passando da un' ottava alla successiva (cioè dal tasto del nostro la a quello del la successivo, posto 8 tasti bianchi

dopo, in quanto le 7 note sono associate ai tasti bianchi) la frequenza associata raddoppia, ci si può chiedere che relazione c'è fra la frequenza fra un tasto e quello successivo. Qui la risposta richiederebbe una conferenza (infatti se ne è tenuta una recentemente nella nostra Scuola su questo argomento). Diciamo solo che se lo strumento è accordato secondo il temperamento "equabile", come quasi sempre accade negli strumenti moderni, il rapporto fra le frequenze associate a due tasti successivi (bianchi o neri) è costante. Dato che fra due note uguali di due ottave successive ci sono 12 tasti (7 bianchi e 5 neri), ne deriva che il rapporto deve essere $^{12}\sqrt{2} \approx 1.06$.



- Rappresentazione ortonormale dei segnali (argomento avanzato, facoltativo)
- Basi ortonormali e sviluppo in serie di Fourier (Argomento avanzato, facoltativo)

▼ 2.2 - Trasformata di Fourier

Trasformata ed integrale di Fourier (funzioni aperiodiche tempo-continue)

La trattazione vista in precedenza per i segnali periodici può essere estesa, con le opportune variazioni, ai segnali aperiodici. Consideriamo una funzione $x(t)$, in generale complessa, aperiodica, definita su tutto l'asse t . Sotto condizioni analoghe a quelle viste in precedenza per le serie di Fourier, esiste la trasformata di Fourier (formula di analisi, dal dominio dei tempi a quello delle frequenze):

$$X(\omega) = \int_{-\infty}^{+\infty} x(t)e^{-i\omega t} dt$$

La cui formula di antitrasformazione è data dalla seguente (formula di sintesi, cioè dalle frequenze ai tempi, qui non dimostrata) è:

$$x(t) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} X(\omega) e^{i\omega t} d\omega$$

Notiamo che l'ampiezza di un fasore ad una determinata pulsazione è infinitesima. Di conseguenza lo spettro di ampiezza bilatero, o meglio la densità spettrale di ampiezza,

$|X(\omega)|$ non è più a righe ma continuo. Analogamente è continuo per lo spettro di fase bilatero, definito come
 $\arg\{X(\omega)\}$.

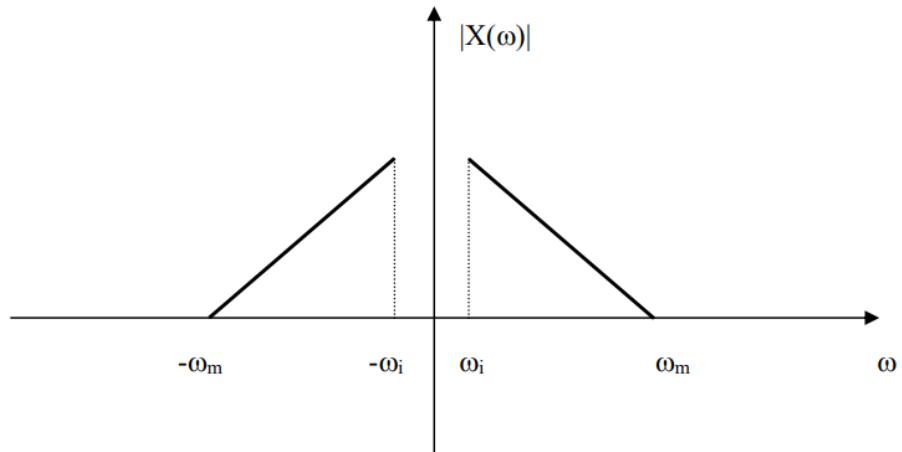


Fig.8 Spettro di ampiezza bilatero (modulo trasformata).

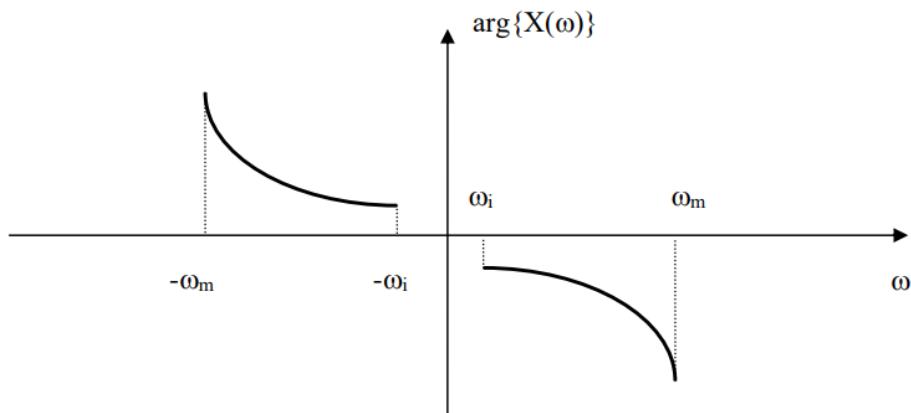


Fig.9 Spettro di fase bilatero (argomento trasformata).

Rappresentazioni monolatere (segnali reali)

Nel caso di un segnale reale si possono dare delle espressioni alternative, monolatere, alla formula di sintesi. Quando $x(t)$ reale, vale la relazione (simmetria hermitiana):

$$X(-\omega) = X^*(\omega)$$

Procedendo in modo analogo a quanto fatto in precedenza per le serie è possibile arrivare alla seguente:

$$\begin{aligned} x(t) &= \frac{1}{2\pi} \int_0^{+\infty} 2\operatorname{Re}\{|X(\omega)|e^{i \arg\{X(\omega)\}} e^{i\omega t}\} d\omega \\ &= \int_0^{+\infty} \frac{|X(\omega)|}{\pi} \cos(\arg\{X(\omega)\} + \omega t) d\omega \end{aligned}$$

Da cui definito lo spettro (densità spettrale) di ampiezza monolatero come:

$$V(\omega) = \frac{|X(\omega)|}{\pi} \quad \omega \geq 0$$

e quello monolatero di fase come:

$$\varphi(\omega) = -\arg\{X(\omega)\} \quad \omega \geq 0, V(\omega) \neq 0$$

Otteniamo infine l'espressione sotto, detta "integrale di Fourier":

$$x(t) = \int_0^{+\infty} V(\omega) \cos[\omega t - \varphi(\omega)] d\omega$$

L' integrale di Fourier è l'analogo dello sviluppo in serie in soli coseni; esso rappresenta $x(t)$ come "somma" di infiniti termini sinusoidali, il generico dei quali ha pulsazione ω , ampiezza (infinitesima) $V(\omega)d\omega$ e fase $\varphi(\omega)$.

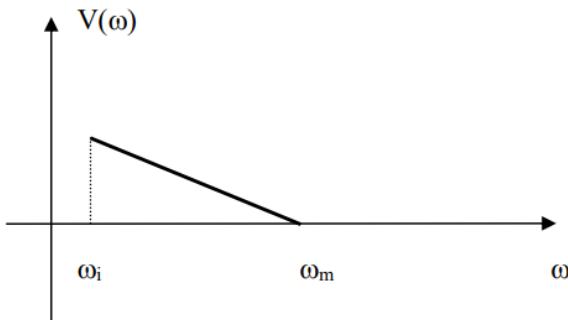


Fig.10 Spettro di ampiezza monolatero.

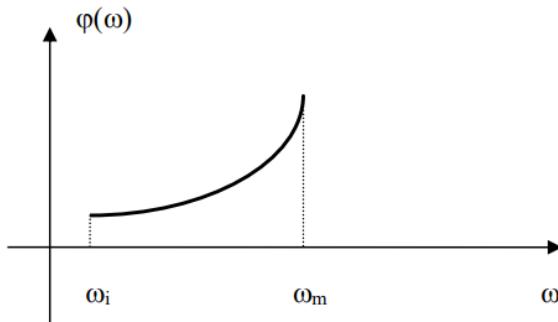


Fig.11 Spettro di fase monolatero.

Bande

Lo spettro di ampiezza consente di introdurre il concetto di banda di pulsazioni di una funzione. Essa è l'intervallo di pulsazioni (o frequenze) per le quali è diverso da zero lo spettro di ampiezza, eventualmente entro un determinato livello di approssimazione. Per comodità, nel caso di segnali reali si considera sempre solo il semiasse positivo anche quando si utilizza uno spettro bilatero. In casi particolari la banda può essere formata da più intervalli disgiunti, ma di norma lo spettro è su di un unico intervallo (almeno approssimativamente).

Quando la banda comprende lo 0 oppure è significativamente prossima ad esso, si parla di funzione passa-basso.

Quando invece la banda $B_\omega = \omega_m - \omega_i$ è piccola rispetto alla pulsazione centrale $\omega_c = \frac{\omega_m + \omega_i}{2}$, cioè tale che risulti $B_\omega \ll \omega_c$ (nel grafico lo spettro di ampiezza è molto lontano dall'origine) si parla di funzione passa-banda.

Proprietà della trasformata di Fourier

Sia $x(t)$ un segnale complesso con trasformata $F[x(t)] = X(\omega)$, valgono le seguenti proprietà ($\xi = t - t_0$):

- Traslazione temporale

$$F[x(t - t_0)] = X(\omega)e^{-j\omega t_0}$$

▼ Dimostrazione

$$\begin{aligned} F[x(t - t_0)] &= \int_{-\infty}^{+\infty} x(t - t_0) e^{-j\omega t} dt = \int_{-\infty - t_0}^{+\infty - t_0} x(\xi) e^{-j\omega(\xi + t_0)} d\xi \\ &= e^{-j\omega t_0} \int_{-\infty}^{+\infty} x(\xi) e^{-j\omega \xi} d\xi = X(\omega) e^{-j\omega t_0} \end{aligned}$$

- Derivata

$$F[\dot{x}(t)] = j\omega X(\omega)$$

▼ Dimostrazione

Scrivendo $x(t)$ come antitrasformata della sua trasformata e derivando si ha

$$\begin{aligned} \dot{x}(t) &= \frac{d}{dt} \frac{1}{2\pi} \int_{-\infty}^{+\infty} X(\omega) e^{j\omega t} d\omega = \frac{1}{2\pi} \int_{-\infty}^{+\infty} X(\omega) \frac{d}{dt} e^{j\omega t} d\omega \\ &= \frac{1}{2\pi} \int_{-\infty}^{+\infty} X(\omega) j\omega e^{j\omega t} d\omega = F^{-1}[j\omega F(\omega)] \end{aligned}$$

Ricaviamo dunque che

$$F[\dot{x}(t)] = F[F^{-1}[j\omega F(\omega)]] = j\omega F(\omega)$$

- Integrale

$$F\left[\int_{-\infty}^t x(\xi) d\xi\right] = \frac{X(\omega)}{j\omega} \quad \text{se } X(0) = 0$$

□ Dimostrazione pag. 25

- Convoluzione

$$F[x(t) * y(t)] = X(\omega)Y(\omega)$$

dove

$$x(t) * y(t) = \int_{-\infty}^{+\infty} x(\tau) y(t - \tau) d\tau$$

□ Dimostrazione pag. 25

È possibile infine dimostrare una relazione tra i coefficienti c_n e la trasformata $X(\omega)$

$$c_n = \frac{1}{T} X(n\omega_0)$$

Esercizi pag. 25 - 47

▼ 3.0 - Trasformata al limite

▼ 3.1 - La funzione generalizzata delta di Dirac

Funzioni, funzionali e distribuzioni

Prima di definire la delta di Dirac conviene ricordare le seguenti definizioni:

- funzione ordinaria: associa ad un numero un altro numero ed uno soltanto (es. funzione coseno)
- funzionale: associa ad una funzione un numero (es. valore medio)
- operatore associa ad una funzione un'altra funzione (es. l'operatore derivata).

Ciò premesso le distribuzioni, o funzioni generalizzate, sono dei funzionali dotati delle proprietà di essere lineari e continui, operanti su delle funzioni dotate di particolari proprietà di regolarità dette funzioni di prova. Il numero che una distribuzione T associa alla funzione di prova φ è indicato con $\langle \varphi, T \rangle$.

Distribuzione delta di Dirac

Data una funzione $x(t)$ continua in $t=0$ (funzione di prova), la distribuzione δ associa ad essa il valore $x(0)$, ovvero campiona la funzione $x(t)$ nell'origine:

$$\langle x, \delta \rangle = x(0)$$

Analogamente si può definire la distribuzione δ_{t_o} come la distribuzione che associa a $x(t)$, continua in t_o , il valore $x(t_o)$, ovvero che campiona la funzione in t_o :

$$\langle x, \delta_{t_o} \rangle = x(t_o)$$

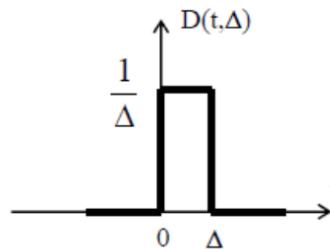
Valutazione come limite in senso integrale

Il valore $x(0)$ di una funzione $x(t)$ continua nell'origine può essere ricavato anche facendo ricorso a famiglie di funzioni ausiliarie $f_\Delta(t)$ per le quali vale

la seguente espressione:

$$x(0) = \lim_{\Delta \rightarrow 0} \int_{-\infty}^{+\infty} x(t) f_{\Delta}(t) dt$$

Fra le famiglie di funzioni ausiliarie $f_{\Delta}(t)$ la più semplice è l'impulso rettangolare di durata Δ ed ampiezza $\frac{1}{\Delta}$ (intensità unitaria), nel seguito indicato con $D(t, \Delta)$.



È possibile inoltre effettuare la seguente sostituzione:

$$\delta(t) = \lim_{\Delta \rightarrow 0} f_{\Delta}(t)$$

da interpretare non in senso ordinario, ma in senso integrale, ottenendo quindi:

$$\langle x, \delta \rangle = x(0) = \lim_{\Delta \rightarrow 0} \int_{-\infty}^{+\infty} x(t) f_{\Delta}(t) dt = \int_{-\infty}^{+\infty} x(t) \delta(t) dt$$

Quanto detto si estende ovviamente anche alle distribuzioni delta ritardate, per le quali:

$$\langle x, \delta_{t_o} \rangle = x(t_o) = \lim_{\Delta \rightarrow 0} \int_{-\infty}^{+\infty} x(t) f_{\Delta}(t - t_o) dt = \int_{-\infty}^{+\infty} x(t) \delta(t - t_o) dt$$

Proprietà

Parità

$$\int_{-\infty}^{+\infty} x(t) \delta(t - t_o) dt = \int_{-\infty}^{+\infty} x(t) \delta(t_o - t) dt$$

Dimostrazione pag. 49

Convoluzione

La funzione generalizzata delta rappresenta l'elemento neutro dell'integrale di convoluzione:

$$x(t) = \int_{-\infty}^{+\infty} x(\tau) \delta(t - \tau) d\tau = x(t) * \delta(t)$$

Dimostrazione pag. 49

Cambio di argomento

$$\delta(\alpha t) = \frac{\delta(t)}{|\alpha|} \quad \alpha \neq 0$$

Dimostrazione pag. 50

Trasformata di Fourier

La trasformata di una delta nell'origine dei tempi corrisponde ad una costante nel dominio delle frequenze:

$$\int_{-\infty}^{+\infty} \delta(t) e^{-j\omega t} dt = 1$$

Dimostrazione pag. 50

Gradino unitario

$$U(t) = \int_{-\infty}^t \delta(\tau) d\tau$$

dove $U(t) = \begin{cases} 1 & t > 0 \\ 0 & t < 0 \end{cases}$.

Dimostrazione pag. 51

▼ 3.2 - Trasformata di Fourier di funzioni periodiche

La trasformata di Fourier di funzioni periodiche non esiste nell'ambito delle funzioni ordinarie, mentre esiste lo sviluppo in serie di Fourier. Sarebbe tuttavia utile, per non dovere duplicare ogni dimostrazione, poter definire una trasformata di Fourier anche per le funzioni periodiche. Ciò è possibile utilizzando le distribuzioni, o funzioni generalizzate, ed in particolare la distribuzione δ .

Dato che in letteratura vengono utilizzate sia le rappresentazioni nelle pulsazioni che nelle frequenze, conviene avere un minimo di familiarità con

entrambe. A questo scopo si ricordano le definizioni di trasformata nelle frequenze (indicata qui con un pedice f) ed antitrasformata:

$$X_f(f) = \int_{-\infty}^{+\infty} x(t)e^{-i2\pi ft} dt$$

$$x(t) = \int_{-\infty}^{+\infty} X_f(f)e^{i2\pi ft} dt$$

Vale la seguente relazione fra trasformate nelle frequenze e trasformate nelle pulsazioni:

$$X_f(f) = X(2\pi f) = X(\omega)$$

Le formule nelle frequenze sono usatissime, in particolare nei testi di Comunicazioni Elettriche, per cui è necessario conoscere anche esse (basta ricordarsi la mancanza della costante).

Trasformate elementari

La trasformata di una costante nei tempi è una delta nell'origine delle frequenze. La trasformata di un esponenziale nei tempi è una delta ritardata. Utilizzando le frequenze si ha:

$$F_f[1] = \delta(f)$$

$$F_f[e^{i2\pi f_o t}] = \delta(f - f_o)$$

Allo stesso modo utilizzando le pulsazioni si ha:

$$F[1] = 2\pi\delta(\omega)$$

$$F[e^{i\omega_o t}] = 2\pi\delta(\omega - \omega_o)$$

Dimostrazione pag. 52

Dalla conoscenza delle trasformate elementari si possono ricavare immediatamente le trasformate dei segnali periodici, per cui ci limitiamo a riportare i risultati.

Trasformata di un segnale sviluppabile in serie di Fourier

$$x(t) = \sum_{n=-\infty}^{+\infty} c_n e^{in\omega_o t}$$

$$\implies X(\omega) = \sum_{n=-\infty}^{+\infty} c_n 2\pi \delta(\omega - n\omega_o)$$

$$X_f(f) = \sum_{n=-\infty}^{+\infty} c_n \delta(f - n f_o)$$

Trasformata del coseno

$$x(t) = \cos \omega_o t = \frac{e^{i\omega_o t} + e^{-i\omega_o t}}{2}$$

$$\implies X(\omega) = \frac{1}{2} \delta(\omega - \omega_o) 2\pi + \frac{1}{2} \delta(\omega + \omega_o) 2\pi$$

$$X_f(f) = \frac{1}{2} \delta(f - f_o) + \frac{1}{2} \delta(f + f_o)$$

Trasformata del seno

$$\sin \omega_o t = \frac{e^{i\omega_o t} - e^{-i\omega_o t}}{2i}$$

$$\implies X(\omega) = \frac{1}{2i} \delta(\omega - \omega_o) 2\pi + \frac{1}{2i} \delta(\omega + \omega_o) 2\pi$$

$$= -i \frac{1}{2} \delta(\omega - \omega_o) 2\pi + i \frac{1}{2} \delta(\omega + \omega_o) 2\pi$$

$$X_f(f) = -i \frac{1}{2} \delta(f - f_o) + i \frac{1}{2} \delta(f + f_o)$$

Trasformata delle funzioni gradino

A partire dalla trasformata del gradino $1(t)$ (formula non dimostrata):

$$F[1(t)] = \frac{1}{i\omega}$$

Dalla definizione di $U(t)$ e $sign(t)$ segue:

$$F[U(t)] = \frac{1}{i\omega} + \frac{2\pi\delta(\omega)}{2}$$

$$F[sign(t)] = \frac{2}{i\omega}$$

Analogamente per le trasformate nelle frequenze si ha (basta sostituire come al solito $\delta(\omega) = \frac{\delta(f)}{2\pi}$).

$$\begin{aligned} F_f[1(t)] &= \frac{1}{i2\pi f} \\ F_f[U(t)] &= \frac{1}{i2\pi f} + \frac{\delta(f)}{2} \\ F_f[sign(t)] &= \frac{1}{i\pi f} \end{aligned}$$

Trasformata di Fourier di un integrale

Poichè

$$y(t) = \int_{-\infty}^t x(\tau)d\tau = \int_{-\infty}^{+\infty} x(\tau)U(t - \tau)d\tau = x(t) * U(t)$$

Trasformando il prodotto di convoluzione si ottiene l'espressione generale della trasformata di un integrale:

$$Y(\omega) = X(\omega)F[U(t)] = \frac{X(\omega)}{i\omega} + \frac{2\pi\delta(\omega)X(\omega)}{2} = \frac{X(\omega)}{i\omega} + \frac{2\pi\delta(\omega)X(0)}{2}$$

Analogamente, per le trasformate in frequenza si ha:

$$Y_f(f) = \frac{X_f(f)}{i2\pi f} + \frac{\delta(f)X_f(0)}{2}$$

▼ 4.0 - Analisi di Fourier di segnali tempo-discreti

▼ 4.1 - Serie temporali

Serie temporali

Le funzioni tempo discrete sono anche chiamate serie temporali (se formate da un numero infinito di termini) o anche successioni o sequenze temporali. Esse possono rappresentare segnali che hanno già origine in tale forma oppure essere ottenute da una funzione tempo continua mediante lettura dei valori da essa assunti in istanti che si succedono con un intervallo T. Questa operazione è detta campionamento e la corrispondente funzione tempo discreta è anche chiamata funzione campionata.

Consideriamo ora la serie temporale:

$$\{x_n\} = \{\dots, x_{-2}, x_{-1}, x_0, x_1, x_2, \dots\}$$

i cui elementi si succedono con intervallo T sull'asse dei tempi e possono essere sia continui che discreti nei valori.

Trasformata di Fourier di una serie

In modo analogo a quanto fatto per le funzioni tempo continue, sotto opportune condizioni, si può definire la trasformata di Fourier della serie temporale mediante la seguente relazione:

$$X_s(\omega) = \sum_{n=-\infty}^{+\infty} x_n e^{-in\omega T}$$

Si noti il pedice s, che serve ad indicare che la trasformata in questione è trasformata di una serie e non di una funzione tempo continua. La formula di antitrasformazione, che dalla trasformata $X_s(\omega)$ permette di tornare agli elementi x_n della serie è la seguente:

$$x_n = \frac{T}{2\pi} \int_{-\frac{\pi}{T}}^{+\frac{\pi}{T}} X_s(\omega) e^{in\omega T} d\omega \quad n = \dots -2, -1, 0, 1, 2, \dots$$

▼ Dimostrazione

Occorre notare che la trasformata $X_s(\omega)$ è funzione periodica con periodo $\omega_P = 2\pi/T$:

$$X_s(\omega) = X_s(\omega + \omega_P)$$

Per questo motivo è possibile esprimere tramite uno sviluppo in serie di Fourier di tipo esponenziale

$$X_s(\omega) = \sum_{n=-\infty}^{\infty} c_n e^{jn\frac{2\pi}{\omega_p}\omega} = \sum_{n=-\infty}^{\infty} c_n e^{jn\frac{2\pi}{T}\omega} = \sum_{n=-\infty}^{\infty} c_n e^{jn\omega T}$$

dove, per definizione di c_n

$$c_n = \frac{1}{\omega_p} \int_{-\frac{\omega_p}{2}}^{+\frac{\omega_p}{2}} X_s(\omega) e^{-jn\frac{2\pi}{\omega_p}\omega} d\omega = \frac{T}{2\pi} \int_{-\frac{\pi}{T}}^{\frac{\pi}{T}} X_s(\omega) e^{-jn\omega T} d\omega$$

Dalla formula della trasformata di Fourier di una serie ricaviamo quindi la relazione

$$x_n = c_{-n}$$

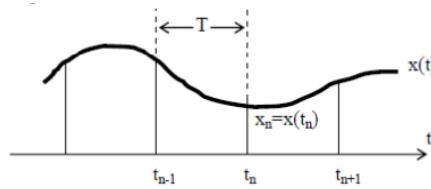
dunque

$$x_n = \frac{T}{2\pi} \int_{-\frac{\pi}{T}}^{+\frac{\pi}{T}} X_s(\omega) e^{in\omega T} d\omega$$

QED.

Serie temporali ottenute per campionamento

L'operazione di campionamento consiste nel leggere i valori di una funzione tempo continua ad intervalli regolari.



I valori $x_n = x(t_n)$ sono detti valori campionati, gli istanti di lettura t_n istanti di campionamento, l'intervallo costante T che li separa è detto intervallo di campionamento e il suo inverso, $\frac{1}{T} = f_o$, frequenza di campionamento.

Senza ledere la generalità possiamo porre $t_n = nT$ (a tal fine basta assumere l'origine dei tempi in uno degli istanti di campionamento) ottenendo la seguente serie temporale:

$$x_n = x(nT) \quad n = \dots, -2, -1, 0, 1, 2, \dots$$

Fra la trasformata $X_s(\omega)$ della serie e la trasformata $X(\omega)$ della funzione campionata vale la seguente importantissima relazione, che lega la trasformata della serie alla ripetizione periodica della trasformata del segnale:

$$X_s(\omega) = \frac{1}{T} \sum_{k=-\infty}^{\infty} X(\omega + k\omega_o)$$

▼ Dimostrazione

In quanto elementi di una serie gli x_n possono essere espressi come antitrasformata della trasformata della serie:

$$x_n = \frac{T}{2\pi} \int_{-\frac{\pi}{T}}^{+\frac{\pi}{T}} X_s(\omega) e^{in\omega T} d\omega$$

In quanto valori di una funzione tempo continua, gli x_n possono essere espressi anche come antitrasformata della trasformata della funzione

tempo continua, calcolata al tempo nT

$$x(nT) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} X(\omega) e^{jn\omega T} d\omega$$

Ora procediamo cercando di rendere il più simile possibile la seconda formula con la prima. Iniziamo spezzando l'intervallo di integrazione in infiniti intervalli, ognuno dei quali di ampiezza pari all'intervallo di integrazione della prima formula, cioè $\omega_o = \frac{2\pi}{T}$

$$x(nT) = \frac{1}{2\pi} \sum_{k=-\infty}^{\infty} \int_{-\frac{\pi}{T} + k\omega_o}^{+\frac{\pi}{T} + k\omega_o} X(\omega) e^{jn\omega T} d\omega$$

Si effettua quindi per ognuno degli infiniti integrali il cambiamento di variabile $\xi = \omega - k\omega_o$ ottenendo una somma di integrali tutti definiti sullo stesso intervallo di integrazione

$$\begin{aligned} x(nT) &= \frac{1}{2\pi} \sum_{k=-\infty}^{\infty} \int_{-\frac{\pi}{T}}^{+\frac{\pi}{T}} X(\xi + k\omega_o) e^{j\xi nT} e^{jk\omega_o nT} d\xi \\ &= \frac{1}{2\pi} \sum_{k=-\infty}^{\infty} \int_{-\frac{\pi}{T}}^{+\frac{\pi}{T}} X(\xi + k\omega_o) e^{j\xi nT} d\xi \end{aligned}$$

Infine, chiamando nuovamente ω la variabile di integrazione, e portando la sommatoria all'interno dell'integrale scriviamo

$$x(nT) = \frac{1}{2\pi} \int_{-\frac{\pi}{T}}^{+\frac{\pi}{T}} \sum_{k=-\infty}^{\infty} X(\omega + k\omega_o) e^{j\omega nT} d\omega$$

Notiamo dunque, con il confronto degli x_n espressi come antitrasformata della trasformata della serie, ovvero

$$x_n = \frac{T}{2\pi} \int_{-\frac{\pi}{T}}^{+\frac{\pi}{T}} X_s(\omega) e^{jn\omega T} d\omega$$

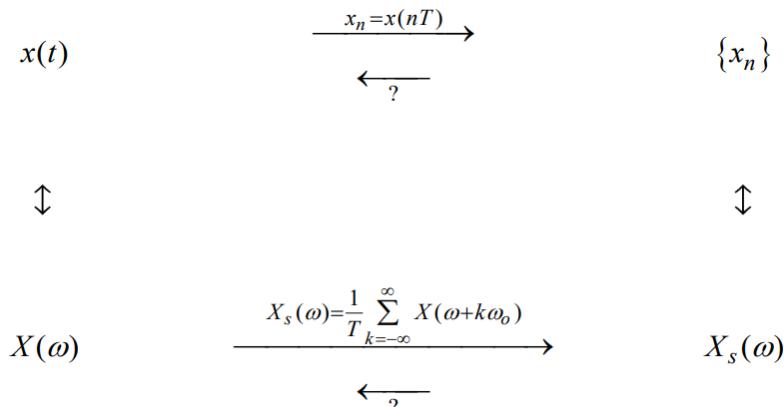
che

$$X_s(\omega) = \frac{1}{T} \sum_{k=-\infty}^{\infty} X(\omega + k\omega_o)$$

QED.

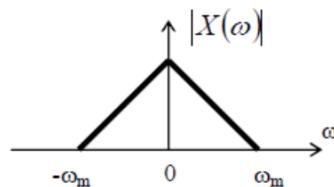
Teorema del campionamento nel dominio dei tempi

Il campionamento trasforma un segnale tempo continuo in un segnale tempo discreto, cioè una serie temporale. Poniamoci ora il problema se e quando la conoscenza dei soli valori campionati è equivalente a quella dell'intero segnale tempo continuo, cioè se e quando il campionamento è reversibile.

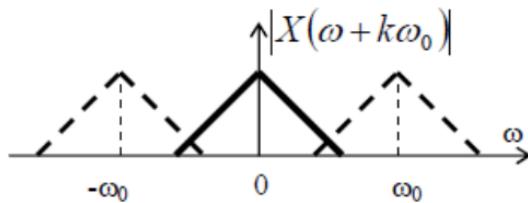


Come si vede dalla figura, se fosse possibile passare dalla ripetizione periodica a $X(\omega)$, da questa si potrebbe risalire a $x(t)$ antitrasformando. La domanda quindi diventa: è possibile risalire dalla ripetizione ad $X(\omega)$? In generale no, ma, se i termini sono non sovrapposti certamente sì.

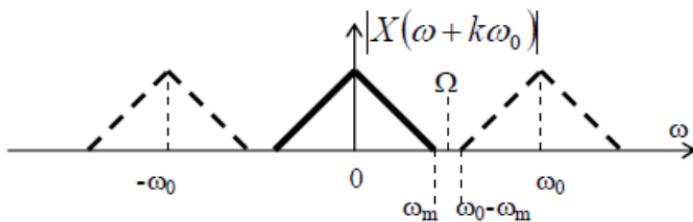
Per procedere consideriamo il caso di una funzione passa-basso, con la massima pulsazione a cui è apprezzabile lo spettro di $x(t)$ pari a ω_m . Un generico andamento di $|X(\omega)|$, in modulo, è dato nella figura sotto (l'andamento scelto, di tipo triangolare, è del tutto arbitrario).



Considerando i moduli dei singoli termini della ripetizione periodica (non il modulo della ripetizione periodica!) si ha la figura sotto.



Come tale figura pone in evidenza, la conoscenza della ripetizione periodica della trasformata del segnale non consente in generale di risalire alla trasformata stessa, e quindi a $x(t)$, in quanto i termini della ripetizione sono sovrapposti, ovvero che si è in presenza di aliasing nel dominio delle frequenze. Aumentando però la frequenza di campionamento, i termini della ripetizione si allontanano, fino a non sovrapporsi più quando $\omega_o > 2\omega_m$, o equivalentemente quando $f_o > 2f_m$ nelle frequenze (assenza di aliasing). Si noti che in questo caso il modulo della ripetizione coincide con la somma dei moduli dei singoli termini, non essendo questi più sovrapposti.



In conclusione, l'enunciato del teorema di campionamento nei tempi è il seguente: dato un segnale passa-basso, condizione sufficiente perché la conoscenza dei valori campionati sia equivalente alla conoscenza della funzione campionata $x(t)$ è che la frequenza di campionamento sia maggiore del doppio della massima frequenza di $x(t)$:

$$\omega_o > 2\omega_m \implies f_o > 2f_m$$

La metà della frequenza di campionamento è detta frequenza di Nyquist, e rappresenta il limite superiore delle frequenze di un segnale passa-basso rappresentabili con una determinata frequenza di campionamento.

Il teorema di campionamento può essere esteso a segnali passa-banda. In questo caso, nella condizione è sufficiente sostituire alla frequenza massima del segnale la sua larghezza di banda.

Sviluppo in serie di Shannon

Sotto la condizione sufficiente del teorema di Shannon, $\omega_o > 2\omega_m$, vogliamo ora esprimere $x(t)$ in funzione dei suoi valori campionati $x(nT)$:

$$x(t) = \sum_{n=-\infty}^{\infty} x_n \frac{\sin \frac{\pi}{T}(t - nT)}{\frac{\pi}{T}(t - nT)} = \sum_{n=-\infty}^{\infty} x_n \text{sinc}\left(\frac{t - nT}{T}\right)$$

La formula sopra prende il nome di sviluppo in serie di Shannon o Whittaker-Shannon, e vale solo se è rispettata la condizione $\omega_o > 2\omega_m$.

Ricordando l'esercizio sull'antitrasformata dell'impulso nelle frequenze si può ricavare che le funzioni $\text{sinc}(\frac{t-nT}{T})$ hanno banda $[0, \frac{\omega_o}{2}]$. Inoltre è possibile dimostrare che esse sono ortogonali. Quindi lo sviluppo in serie di Shannon rappresenta uno sviluppo di $x(t)$ in serie di funzioni ortogonali i cui coefficienti sono gli stessi valori campionati $x(nT)$. Tale interpretazione è resa possibile dalla scelta della metà della pulsazione di campionamento per isolare il termine centrale.

▼ Dimostrazione

Siccome la trasformata della serie equivale alla ripetizione periodica della trasformata del segnale, a meno di una costante moltiplicativa, isolando il termine centrale della ripetizione otteniamo

$$X(\omega) = \begin{cases} TX_s(\omega) & |\omega| < \omega_o/2 \\ 0 & \text{altrove} \end{cases}$$

Inseriamo dunque la formula sopra nella formula di antitrasformazione del segnale

$$x(t) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} X(\omega) e^{j\omega t} d\omega = \frac{T}{2\pi} \int_{-\frac{\pi}{T}}^{+\frac{\pi}{T}} X_s(\omega) e^{j\omega t} d\omega$$

Sostituendo dunque la definizione di $X_s(\omega)$ nella formula sopra otteniamo dunque

$$x(t) = \frac{T}{2\pi} \sum_{n=-\infty}^{\infty} x_n \int_{-\frac{\pi}{T}}^{+\frac{\pi}{T}} e^{j\omega t} e^{-jn\omega T} d\omega$$

risolvendo parte della quale (passaggi visti negli esercizi) si arriva alla formula

$$x(t) = \sum_{n=-\infty}^{\infty} x_n \frac{\sin \frac{\pi}{T}(t - nT)}{\frac{\pi}{T}(t - nT)} = \sum_{n=-\infty}^{\infty} x_n \text{sinc}\left(\frac{t - nT}{T}\right)$$

QED.

Proprietà della trasformata di una serie temporale

Serie ritardata

Se $F[\{x_n\}] = X_s(\omega)$, allora

$$F[\{x_{n-m}\}] = X_s(\omega)e^{-i\omega mT}$$

Si noti che la formula è la stessa delle funzioni tempo continue dato che a m posizioni corrisponde il ritardo mT.

Convoluzione fra serie temporali

Date due serie temporali $\{x_n\}$ e $\{y_n\}$ la loro convoluzione definisce una nuova serie temporale $\{z_n\}$ i cui termini sono espressi da:

$$z_n = x_n * y_n = \sum_{i=-\infty}^{\infty} x_i y_{n-i}$$

Anche in questo caso l'analogia con il caso tempo continuo è ampia (basta sostituire una sommatoria in i all'integrale in τ).

La trasformata del prodotto di convoluzione è come nel caso continuo data dal prodotto delle trasformate:

$$Z_s(\omega) = X_s(\omega)Y_s(\omega)$$

dove $Z_s(\omega)$ è la trasformata della serie $z_n = x_n * y_n$.

Dimostrazione pag. 60

Convoluzione fra una serie temporale ed una funzione tempo-continua

Data la serie temporale $\{x_n\}$ e la funzione tempo-continua $g(t)$, la loro convoluzione definisce una funzione tempo continua espressa da:

$$y(t) = \{x_n\} * g(t) = \sum_{n=-\infty}^{\infty} x_n g(t - nT)$$

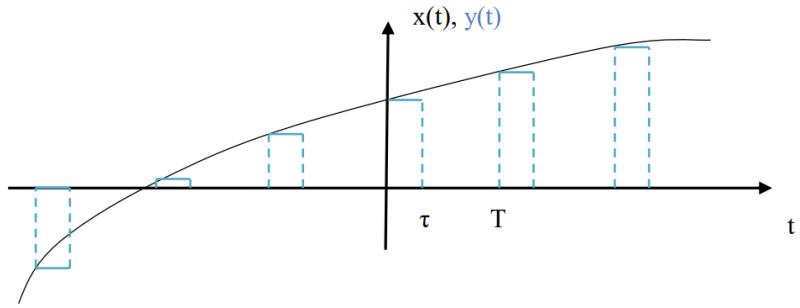
La trasformata del prodotto di convoluzione è al solito il prodotto delle trasformate:

$$Y(\omega) = X_s(\omega)G(\omega)$$

Dimostrazione pag. 60

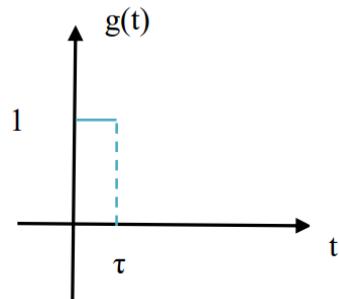
Trasformata di un segnale PAM ottenuto da una serie di campioni

Si consideri la successione di impulsi rettangolari ottenuta campionando una funzione tempocontinua $x(t)$ con intervallo T e mantenendo i valori campionati per un intervallo τ con $\tau < T$.



La successione di impulsi in esame, aventi intervallo di ripetizione e durata costanti, ma ampiezze variabili, costituisce una successione di impulsi modulata in ampiezza e viene denominata segnale PAM (Pulse Amplitude Modulation).

Il segnale PAM in questione può essere visto come convoluzione fra la serie temporale $\{x_n\}$ e l'impulso rettangolare $g(t)$, di ampiezza unitaria e durata τ



Si ha infatti:

$$y(t) = \sum_{n=-\infty}^{\infty} x_n g(t - nT) = \{x_n\} * g(t)$$

Poiché l'operatore trasformata di Fourier trasforma i prodotti di convoluzione in prodotti semplici e ricordando il legame fra trasformata di una serie ottenuta per campionamento e trasformata della funzione campionata si ottiene:

$$Y(\omega) = X_s(\omega)G(\omega) = \frac{1}{T} \sum_{k=-\infty}^{\infty} X(\omega + k\omega_o)G(\omega)$$

Si noti infine che per esprimere la trasformata $G(\omega)$ basta ricordare l'esercizio sulla trasformata di un impulso rettangolare ed osservare che ora

l'impulso $g(t)$ è centrato sull'istante $\tau/2$ anzichè sull'origine, cioè è ritardato di $\tau/2$:

$$G(\omega) = \tau \frac{\sin(\omega\tau/2)}{\omega\tau/2} e^{-i\frac{\omega\tau}{2}}$$

▼ 4.2 - La trasformata di Fourier discreta (DFT)

La trasformata di Fourier discreta (DFT)

Finora abbiamo visto trasformate di Fourier di segnali tempo-continui, quindi di segnali tempo-discreti, ovvero di serie temporali. In entrambi i casi il segnale è generalmente definito su tutto l'asse dei tempi. Consideriamo ora un nuovo tipo di trasformata, detta trasformata di Fourier discreta (DFT=Discrete Fourier Transform), che si applica non più a una serie di infiniti termini ma ad una npla, cioè ad un vettore, costituito in generale da componenti complesse. Più precisamente la DFT stabilisce una corrispondenza biunivoca fra n-ple di numeri, in generale complessi:

$$(x_0, x_1, \dots, x_{N-1}) \xrightleftharpoons{DFT} (X_0, X_1, \dots, X_{N-1})$$

L'elemento q -esimo dell'n-pla di arrivo è definito come (formula di trasformazione):

$$X_q = \sum_{n=0}^{N-1} x_n e^{-i \frac{2\pi}{N} n q}$$

Si noti come questa formula assomigli alla trasformata di una serie temporale, a patto di considerare un numero finito di termini, di sostituire q a ω , e $\frac{2\pi}{N}$ a T .

La formula di anttrasformazione (IDFT=Inverse Discrete Fourier Transform), che ci restituisce un termine della npla di partenza a partire da quella di arrivo è la seguente:

$$x_n = \frac{1}{N} \sum_{q=0}^{N-1} X_q e^{i \frac{2\pi}{N} n q}$$

Anche qui l'analogia è abbastanza stretta. Si tratta come per tutte le anttrasformate di cambiare segno all'argomento dell'esponenziale, quindi di inserire un termine moltiplicativo. Questo era $T/2\pi$ nell'anttrasformata delle

serie temporali, qui è $1/N$. Si noti che per passare dal primo al secondo è sufficiente sostituire $\frac{2\pi}{N}$ a T .

Dimostrazione pag. 62

Legame tra trasformata di Fourier discreta e continua

La trasformata di Fourier discreta costituisce un elemento fondamentale dell'elaborazione digitale dei segnali (Digital Signal Processing). Essa può in particolare essere utilizzata per calcolare, in modo approssimato, la trasformata di Fourier di segnali tempo continui.

Sia dunque $x(t)$ una funzione con trasformata $X(\omega)$.

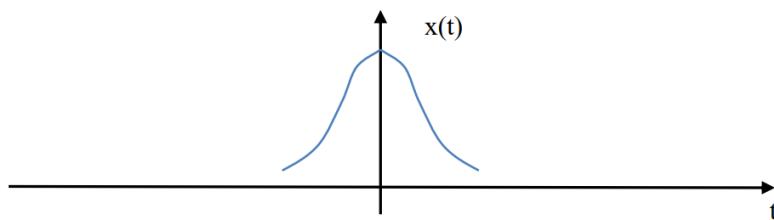


Fig.50 Funzione del tempo (arbitraria)

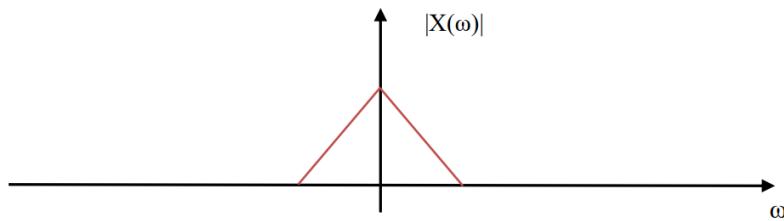


Fig.51 Sua trasformata (modulo)

Risulterebbe comodo che prendendo N campioni di $x(t)$ ed eseguendo la DFT si ottenessero N campioni della trasformata $X(\omega)$. Il legame è tuttavia più complesso, e riguarda le ripetizioni periodiche di $x(t)$ e di $X(\omega)$.

Costruiamo pertanto le due ripetizioni periodiche:

$$x_p(t) = \sum_{i=-\infty}^{+\infty} x(t - iT_p)$$

$$X_p(\omega) = \sum_{i=-\infty}^{+\infty} X(\omega - i\omega_p)$$

con periodi rispettivamente T_p e ω_p legati fra loro dalla relazione $T_p\omega_p = 2\pi N$.

Definiamo quindi gli intervalli nei tempi e nelle frequenze, dividendo i periodi per il numero di punti:

$$\Delta t = \frac{T_p}{N} = \frac{2\pi}{\omega_p}$$

$$\Delta\omega = \frac{\omega_p}{N} = \frac{2\pi}{T_p}$$

E si noti che gli intervalli soddisfano la relazione $\Delta t \Delta \omega = \frac{2\pi}{N}$.

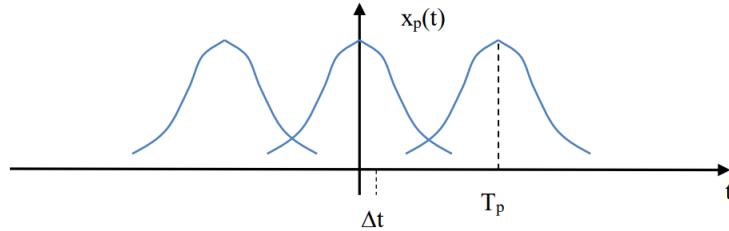


Fig.52 Ripetizione periodica della funzione del tempo (evidenziati i singoli termini).

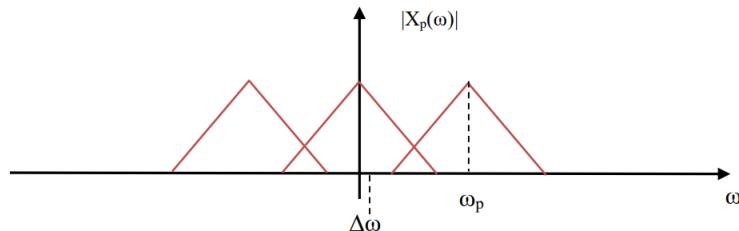


Fig.53 Ripetizione periodica della sua trasformata (evidenziato il modulo dei singoli termini)

Prendiamo quindi N campioni della ripetizione nei tempi a partire dall'origine con intervallo Δt . Si noti che l'ultimo punto di campionamento, $(N - 1)\Delta t$, sarà appena prima del periodo, essendo $T_p = N\Delta t$.

Prendendo quindi N campioni della ripetizione nelle frequenze in modo analogo, si ottengono le seguenti due n-pie ordinate di numeri:

$$x_n = x_p(n\Delta t) \quad n = 0, 1, \dots, N - 1$$

$$X_n = X_p(n\Delta\omega) \quad n = 0, 1, \dots, N - 1$$

Moltiplichiamo quindi per Δt la n-pla nei tempi; si può dimostrare che essa è legata alla n-pla nelle frequenze dalla trasformazione discreta di Fourier:

$$(x_0\Delta t, x_1\Delta t, \dots, x_{N-1}\Delta t) \xrightleftharpoons{DFT} (X_0, X_1, \dots, X_{N-1})$$

Il legame è quindi (a meno della costante Δt) fra N campioni nei tempi ed N campioni nelle frequenze, non di $x(t)$ e $X(\omega)$, ma delle loro ripetizioni periodiche, costruite come sopra indicato (dimostrazione facoltativa).

Quando i termini delle due ripetizioni non si sovrappongono in modo significativo, ovvero è trascurabile l'aliasing sia nel dominio dei tempi che in

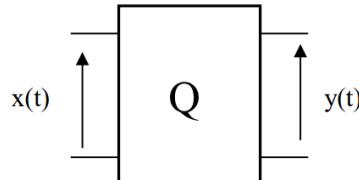
quello delle frequenze, le N-pie nei tempi e nelle frequenze, in generale rappresentative delle ripetizioni periodiche, diventano rappresentative delle sole $x(t)$ e $X(\omega)$. Se ciò si verifica la DFT può essere utilizzata per calcolare la trasformata continua. Per evitare l'aliasing occorre assumere sia T_p e ω_p sufficientemente grandi, il che comporta la scelta un numero di punti N sufficientemente elevato.

▼ 5.0 - Sistemi lineari

▼ 5.1 - Sistemi lineari tempo-continui

Sistemi lineari

Consideriamo il sistema ingresso-uscita rappresentato nella figura sotto, in cui $x(t)$ indica un segnale d'ingresso tempo-continuo e $y(t) = Q[x(t)]$ la corrispondente risposta, essa pure tempo-continua.



La risposta $y(t)$ è una trasformazione dell'ingresso $x(t)$, cioè in generale dipende dall'intero andamento di $x(t)$. Condizione necessaria per la fisica realizzabilità è che il sistema sia causale, cioè che ad un determinato istante l'uscita dipenda dai valori passati e da quello attuale dell'ingresso, ma non dai valori futuri. Il sistema si dice poi algebrico se l'uscita dipende solo dal valore attuale dell'ingresso (il sistema non ha memoria), quindi si ha $y = f(x)$, cioè l'uscita è funzione dell'ingresso. Il sistema si dice lineare se

$$Q[c_1x_1(t) + c_2x_2(t)] = c_1Q[x_1(t)] + c_2Q[x_2(t)]$$

Qualunque siano i coefficienti e i segnali in ingresso.

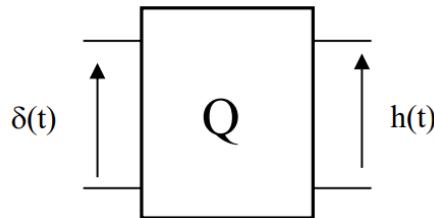
Il sistema è detto tempo-invariante se la risposta al segnale ritardato è la risposta ritardata, qualunque sia il ritardo t_0 :

$$y(t - t_0) = Q[x(t - t_0)]$$

Facciamo ora riferimento a sistemi lineari tempo-invarianti per dare le definizioni di risposta impulsiva e di funzione di trasferimento.

Risposta impulsiva di un sistema lineare

Un sistema lineare tempo invariante può essere completamente caratterizzato nel dominio dei tempi dalla sua risposta impulsiva $h(t)$. Essa è definita come la risposta della rete all'impulso di Dirac $\delta(t)$. La risposta impulsiva $h(t)$ può essere reale o complessa. Qui ci limiteremo al caso reale.



Dato che la Delta di Dirac è una distribuzione e non una funzione ordinaria, può essere comodo dare una definizione alternativa della risposta impulsiva, di carattere più operativo. Si consideri a questo scopo in ingresso la funzione ausiliaria $D(t, \Delta)$, ovvero l'impulso rettangolare di durata Δ ed ampiezza $\frac{1}{\Delta}$.

La risposta $y_\Delta(t)$ alla funzione ausiliaria dipende dal tempo ma anche da Δ . Il suo limite coincide con la risposta impulsiva $h(t)$.

$$h(t) = \lim_{\Delta \rightarrow 0} y_\Delta(t)$$

Per un sistema causale, condizione necessaria per la fisica realizzabilità, è la seguente:

$$h(t) = 0 \quad t < 0$$

La risposta impulsiva consente sempre di esprimere l'uscita della rete quando al suo ingresso è presente un generico segnale $x(t)$, rimanendo nel dominio dei tempi, tramite la relazione:

$$y(t) = x(t) * h(t)$$

Dimostrazione pag. 69

Funzione di trasferimento di una rete lineare

La caratterizzazione nel dominio delle frequenze di una rete lineare è data dalla funzione di trasferimento $H(\omega)$. Essa può essere definita in diversi modi, noi avendo già introdotto la risposta impulsiva la definiamo come trasformata di Fourier della risposta impulsiva:

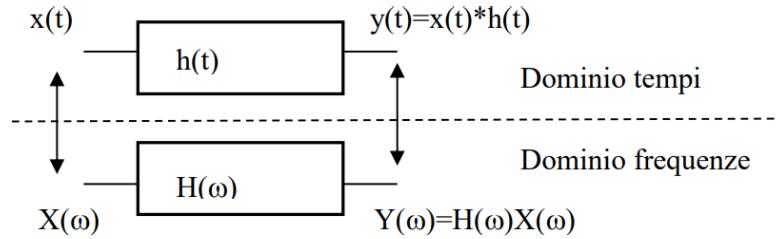
$$H(\omega) = F[h(t)]$$

Data questa definizione, dalla relazione $y(t) = x(t) * h(t)$ è immediato, ricordando il teorema della convoluzione, ricavare la seguente:

$$Y(\omega) = X(\omega)H(\omega)$$

Nel caso in cui $h(t) \in \mathbb{R}$, cioè la rete è reale, vale la simmetria hermitiana

$$H(\omega) = H^*(-\omega)$$



Caratteristiche di ampiezza e fase

La funzione di trasferimento è in generale complessa; il medesimo contenuto informativo può essere dato da due funzioni reali, le caratteristiche di ampiezza e fase, prese insieme:

$$\begin{cases} T(\omega) = |H(\omega)| \\ \beta(\omega) = -\arg\{H(\omega)\} \end{cases}$$

Il segno meno davanti all'argomento dipende dalla convenzione adottata per le fasi.

La simmetria hermitiana della funzione di trasferimento (nel caso $h(t) \in \mathbb{R}$) equivale ad avere caratteristica di ampiezza pari e fase dispari:

$$\begin{cases} T(\omega) = T(-\omega) \\ \beta(\omega) = -\beta(-\omega) \end{cases}$$

Proprietà

Risposta ad un fasore

Ad un fasore in ingresso una rete lineare tempo-invariante risponde con un fasore in uscita, avente medesima frequenza angolare e diverso numero complesso rappresentativo. Se il segnale in ingresso è $x(t) = c_x e^{j\omega_1 t}$, la rete risponde con

$$y(t) = c_y e^{j\omega_1 t}, \quad c_y = c_x H(\omega_1)$$

Dimostrazione pag. 71

Risposta ad una sinusoide

Ad una sinusoide in ingresso una rete lineare tempo-invariante reale risponde con una sinusoide in uscita, avente medesima frequenza angolare e diversa ampiezza e fase, cioè diverso numero complesso rappresentativo. Se $x(t) = A_1 \cos(\omega_1 t - \varphi_x)$, la rete risponde con

$$y(t) = A_y \cos(\omega_1 t - \varphi_y)$$

dove $A_y = A_x T(\omega_1)$ e $\varphi_y = \varphi_x + \beta(\omega_1)$.

Dimostrazione pag. 71

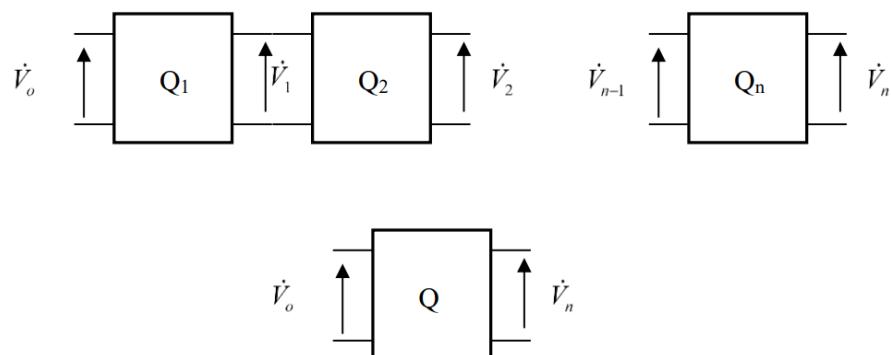
Risposta a segnali sviluppabili in serie di Fourier o trasformabili secondo Fourier

Riassumendo, la conoscenza della funzione di trasferimento permette di ricavare la risposta nel dominio delle frequenze indipendentemente dal tipo di rappresentazione del segnale in ingresso. Si ricordano le seguenti formule specifiche, di immediata derivazione, utili per gli esercizi.

$X(\omega)$ $x(t) = \sum_{n=-\infty}^{\infty} c_n e^{jn\omega_0 t}$ $x(t) = A_o + \sum_{n=1}^{\infty} A_n \cos(n\omega_0 t - \varphi_n)$ $x(t) = \int_0^{+\infty} V(\omega) \cos[\omega t - \varphi(\omega)] d\omega$	$Y(\omega) = X(\omega)H(\omega)$ $y(t) = \sum_{n=-\infty}^{\infty} c_n H(n\omega_0) e^{jn\omega_0 t}$ $y(t) = A_o H(0) + \sum_{n=1}^{\infty} A_n T(n\omega_0) \cos[n\omega_0 t - \varphi_n - \beta(n\omega_0)]$ $y(t) = \int_0^{+\infty} V(\omega) T(\omega) \cos[\omega t - \varphi(\omega) - \beta(\omega)] d\omega$
--	---

Sistemi in cascata

Consideriamo più sistemi lineari in cascata: la funzione di trasferimento della cascata è uguale al prodotto delle funzioni di trasferimento.



Conviene considerare un regime sinusoidale; con riferimento ai numeri complessi rappresentativi delle funzioni in gioco, per la funzione di trasferimento complessiva risulta:

$$H(\omega) = \frac{\dot{V}_n}{\dot{V}_0}$$

E per le funzioni dei vari blocchi:

$$H_1(\omega) = \frac{\dot{V}_1}{\dot{V}_0} \quad H_2(\omega) = \frac{\dot{V}_2}{\dot{V}_1} \dots H_N(\omega) = \frac{\dot{V}_N}{\dot{V}_{N-1}}$$

Da cui è immediato:

$$H(\omega) = H_1(\omega)H_2(\omega) \dots H_N(\omega)$$

Esercizi pag. 72

▼ 5.2 - Filtri

Condizioni di non distorsione

Si dice che il segnale $y(t)$ riproduce indistorto $x(t)$ quando differisce da esso solo per una costante moltiplicativa ed un ritardo:

$$y(t) = Ax(t - t_0)$$

Trasformando questa secondo Fourier si ha:

$$Y(\omega) = AX(\omega)e^{-j\omega t_0}$$

Da cui:

$$H(\omega) = Ae^{-j\omega t_0}$$

Restringendosi al caso di ingresso e rete reali, anche l'uscita sarà reale. In questo caso la costante moltiplicativa A sarà anch'essa reale. Assumendola positiva e chiamandola T_0 , dalla condizione di non distorsione enunciata per $H(\omega)$ si ricavano le equivalenti condizioni di non distorsione per le caratteristiche di ampiezza e fase, avendo indicato con B_x la banda di $x(t)$:

$$\begin{cases} T(\omega) = T_0 & \omega \in B_x \\ \beta(\omega) = \omega t_0 \end{cases}$$

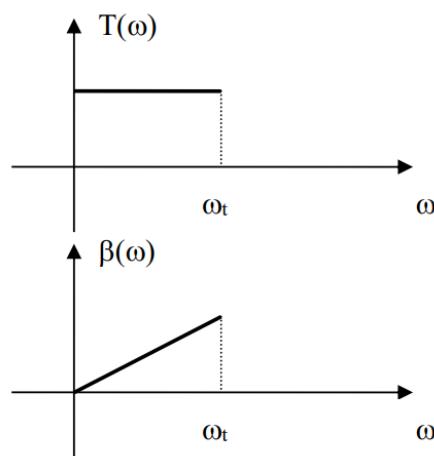
Volendone dare un'interpretazione facile da ricordare, possiamo dire che le condizioni di non distorsione sono condizioni di equità nel dominio delle frequenze rispetto alle componenti dell'ingresso. Si supponga quest'ultimo formato da n sinusoidi. La prima condizione richiede che ognuna di esse sia amplificata/attenuata esattamente dello stesso valore, cioè nessuna di esse deve essere enfatizzata rispetto alle altre. La seconda che tutte siano ritardate dello stesso ritardo (non che escano prima gli alti, poi i bassi, poi i medi...).

Filtri ideali

Strettamente legato alle condizioni di non distorsione è il concetto di filtri ideali, la cui descrizione naturale è nel dominio delle frequenze, cioè in termini di caratteristiche di ampiezza e fase. Un filtro ideale è un sistema lineare avente la proprietà di fare passare alcune componenti in frequenza, senza distorcerle, eliminando le altre. Essi sono caratterizzati da una o più bande passanti, dove valgono le condizioni di non distorsione, e da una o più bande attenuate, dove la caratteristica di ampiezza è nulla. Il passaggio dalle une alle altre avviene alle cosiddette frequenze di taglio.

Passa-basso

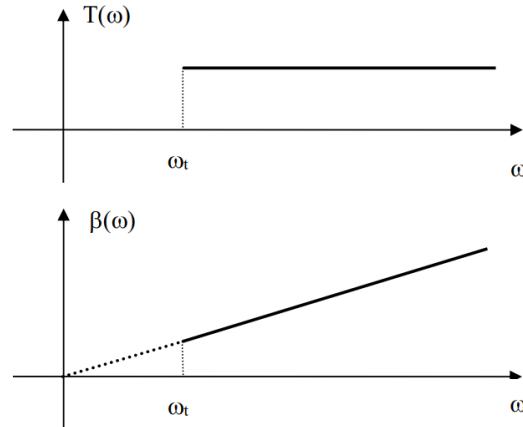
Il filtro in questione lascia passare indistorte tutte le componenti a frequenza inferiore alla frequenza di taglio, eliminando tutte le altre. È caratterizzato da un'unica frequenza di taglio (sul semiasse positivo). Si noti il simbolo, autoesplicativo



Caratteristica di ampiezza e fase di un filtro ideale passa-basso.

Passa-alto

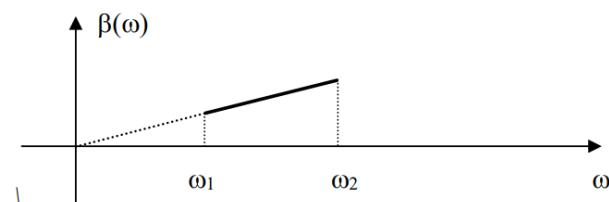
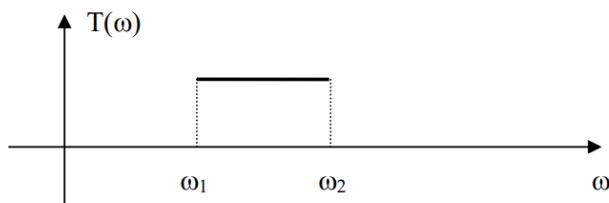
E' il duale del passa-basso.



Caratteristica di ampiezza e fase di un filtro ideale passa-alto.

Passa-banda

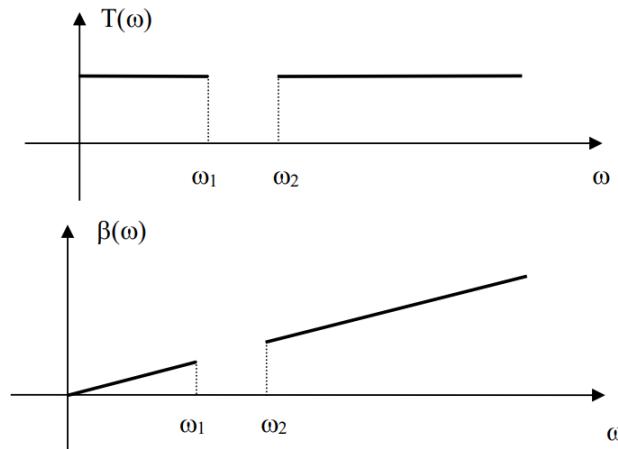
E' equivalente all'applicazione in cascata di un passa-basso e di un passa-alto; è caratterizzato sul semiasse positivo da due frequenze di taglio che delimitano la banda passante.



Caratteristiche di ampiezza e fase di un filtro ideale passa-banda.

Elimina-banda

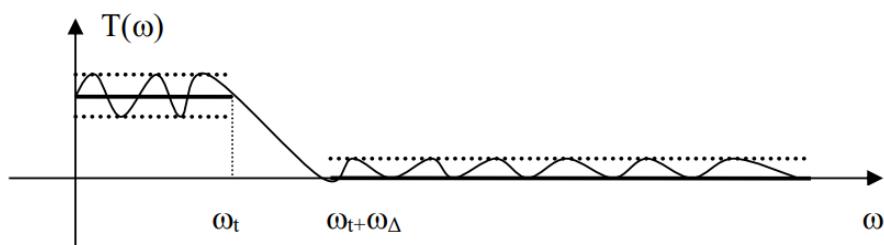
E' il duale del passa-banda.



Caratteristiche di ampiezza e fase di un filtro ideale elimina-banda.

I filtri ideali permettono di isolare il segnale utile da dei disturbi, quando questi ultimi occupano bande diverse. Questa però non è l'unica applicazione dei filtri. Infatti essi si applicano anche per ridurre la banda del segnale utile, in modo da farla rientrare nella banda disponibile per un dato servizio.

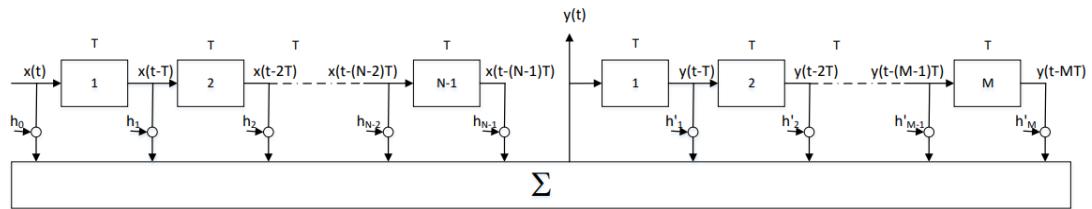
I filtri ideali ora presentati non sono fisicamente realizzabili. Infatti la caratteristica di ampiezza non può mai presentare discontinuità come quelle prima evidenziate in corrispondenza alle pulsazioni di taglio. Ad esempio, la caratteristica di ampiezza di un possibile filtro reale passa-basso è riportata nella figura sotto, dove sono evidenziati i margini di tolleranza in banda passante ed in banda attenuata, e l'intervallo ω_Δ finito fra la due bande per permettere il passaggio graduale, senza la discontinuità propria del filtro ideale, fra le due.



Filtri ricorrenti

Consideriamo lo schema a blocchi del sistema rappresentato nella figura sotto: esso è costituito da due catene di blocchi di ritardo (pari a T), una diretta ed una in retroazione. La prima è formata da $N - 1$ blocchi e da N

prese che prelevano versioni diversamente ritardate del segnali d'ingresso; esse moltiplicate per i coefficienti h_k entrano nel sommatore in basso. Da esso viene prelevata l'uscita $y(t)$ che viene posta in ingresso alla catena di retroazione, simile alla prima ma con M elementi di ritardo ed M prese, da cui vengono prelevate le versioni diversamente ritardate dell'uscita. Una volta moltiplicate per i coefficienti h'_k vengono anche esse inserite nel sommatore.



La relazione che descrive analiticamente quanto detto, è la seguente:

$$y(t) = \sum_{k=0}^{N-1} h_k x(t - kT) + \sum_{k=1}^M h'_k y(t - kT)$$

Da essa, applicando la trasformata di Fourier ad entrambi i membri e ricordando la trasformata di un segnale ritardato si ha

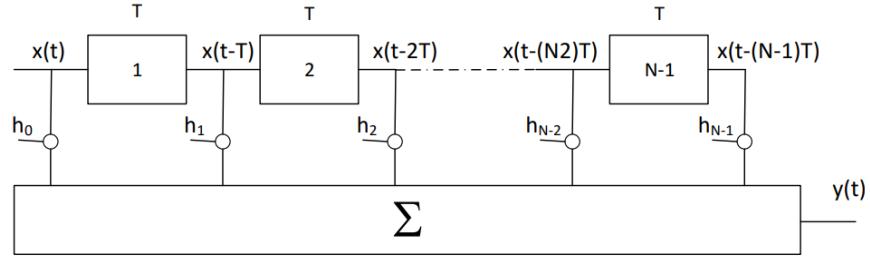
$$\begin{aligned} Y(\omega) &= \sum_{k=0}^{N-1} h_k X(\omega) e^{-j\omega kT} + \sum_{k=1}^M h'_k Y(\omega) e^{-j\omega kT} = \\ &= X(\omega) \sum_{k=0}^{N-1} h_k e^{-j\omega kT} + Y(\omega) \sum_{k=1}^M h'_k e^{-j\omega kT} \end{aligned}$$

Riordinando è immediato ricavare la funzione di trasferimento come rapporto delle trasformate:

$$H(\omega) = \frac{Y(\omega)}{X(\omega)} = \frac{\sum_{k=0}^{N-1} h_k e^{-j\omega kT}}{1 - \sum_{k=1}^M h'_k e^{-j\omega kT}}$$

Filtrati trasversali

Nei filtri trasversali manca il ramo di retroazione, ovvero i coefficienti h'_k sono tutti nulli.



La relativa funzione di trasferimento è:

$$H(\omega) = \sum_{k=0}^{N-1} h_k e^{-i\omega kT}$$

Filtri puramente ricorrenti

In essi manca la catena diretta e l'ingresso vien portato direttamente al sommatore, cioè $h_0 = 1$ mentre sono tutti nulli gli altri coefficienti h_k .

La relativa funzione di trasferimento è:

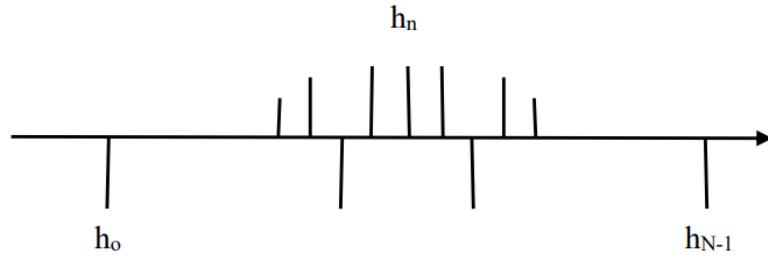
$$H(\omega) = \frac{1}{1 - \sum_{k=1}^M h'_k e^{-i\omega kT}}$$

Criteri di progetto di filtri trasversali a coefficienti simmetrici

Il progetto dei filtri ricorrenti non verrà qui esaminato in dettaglio. Tuttavia, considereremo qui un caso particolare, quello di filtri trasversali a coefficienti simmetrici, impiegati per la realizzazione di un filtro passabasso. Per i motivi che vedremo fra poco imponiamo la simmetria dei coefficienti del filtro trasversale, cioè:

$$h_k = h_{N-1-k}$$

Consideriamo il caso di N dispari.



In questo caso il coefficiente centrale non ha un corrispondente e la condizione di simmetria sopra può essere riscritta in riferimento ad esso. Quindi indicato con $n=(N-1)/2$ il pedice del coefficiente centrale si ha

$$h_{n-r} = h_{n+r}$$

Conviene quindi estrarre dalla funzione di trasferimento l'esponenziale $e^{-i\omega nT}$ e scrivere $N - 1$ come $2n$ per i motivi che vedremo in seguito

$$H(\omega) = e^{-j\omega nT} \sum_{k=0}^{2n} h_k e^{-j\omega(k-n)T}$$

Si cambia indice della sommatoria, ponendo $n = k - r$ per renderla simmetrica; quindi si procede in modo simile a quanto fatto per introdurre la II forma dello sviluppo in serie di Fourier

$$H(\omega) = e^{-j\omega nT} \sum_{r=-n}^n h_{n+r} e^{-j\omega rT} = e^{-j\omega nT} \left(h_n + \sum_{r=1}^n 2h_{n+r} \cos \omega rT \right)$$

Ponendo

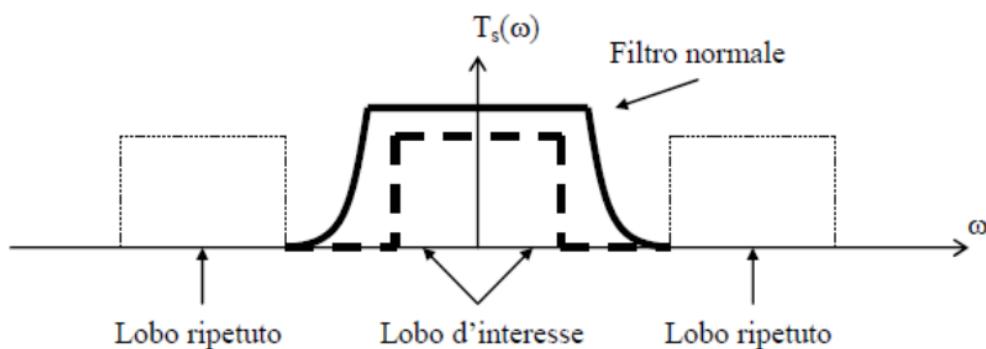
$$G(\omega) = h_n + \sum_{r=1}^n 2h_{n+r} \cos \omega rT$$

si ha infine

$$H(\omega) = G(\omega) e^{-jn\omega T}$$

La funzione $G(\omega)$ è funzione reale, pari e periodica con periodo $2\pi/T$. Il suo modulo rappresenta la caratteristica di ampiezza, mentre la loro

caratteristica di fase, quando $G(\omega) > 0$, è proporzionale a ω . Si noti che questa proporzionalità soddisfa automaticamente la seconda delle condizioni di non distorsione. Il progetto del filtro dovrà fare in modo che $G(\omega)$ sia circa costante nelle bande passanti e circa zero in quelle attenuate. In realtà, a causa della periodicità di $G(\omega)$, le bande passanti desiderate si ripetono con periodo $2\pi/T$, dando origine a bande passanti spurie (si veda a questo proposito l'esempio nella figura sotto, riferito ad un passa-basso). Se l'ingresso del filtro non comprende componenti spettrali nelle bande passanti spurie esse sono ininfluenti; altrimenti occorre porre in cascata al filtro trasversale un filtro supplementare che provveda ad eliminare tali bande.



Procedimento sub-ottimo per la realizzazione di un passa-basso

Il progetto di un filtro trasversale richiede il calcolo dei valori di N e dei coefficienti h_k che rendono soddisfatte, nelle bande di competenza e secondo specifiche assegnate, le approssimazioni $G(\omega) = 1$ nella bande passanti e $G(\omega) = 0$ nelle bande attenuate.

Esistono algoritmi ottimali che, date le specifiche di tolleranza, realizzano l'andamento voluto della $G(\omega)$ con andamenti di tipo oscillatorio attorno ai valori desiderati 1 e 0. Vale però la pena parlare di un procedimento sub-ottimo, che viene suggerito immediatamente dall'espressione della $G(\omega)$. Osserviamo infatti che se fosse $n=\infty$ questa relazione altro non sarebbe che lo sviluppo in serie di Fourier di una funzione periodica e pari (sviluppo in soli coseni, seconda forma). Lo sviluppo in questione ci darebbe quindi i coefficienti cercati dati da:

$$h_n = a_0/2$$

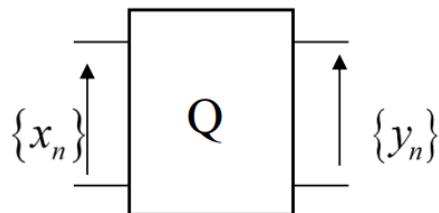
$$h_{n+r} = h_{n-r} = a_r/2$$

Con infiniti termini sarebbe possibile realizzare qualunque caratteristica di ampiezza, anche ideale, con discontinuità in corrispondenza alle pulsazioni

di taglio. Ciò richiederebbe però un filtro di lunghezza infinita, non fisicamente realizzabile, che peraltro introdurrebbe anche un ritardo infinito. Ovviamente occorre quindi limitare il numero dei termini. Un procedimento sub-ottimo per realizzare, nell'ambito di una prefissata approssimazione, una qualunque caratteristica di ampiezza potrebbe allora essere il seguente: dapprima si sviluppa in serie di Fourier la caratteristica assegnata, quindi si provvede poi a troncare tale sviluppo in modo da mantenere un numero di termini sufficiente ad approssimare tale caratteristica secondo le specifiche stabilitate.

▼ 5.3 - Sistemi lineari tempo-discreti

La definizione dei sistemi lineari tempo discreti ricalca quella dei sistemi lineari tempo continui, a meno della simbologia. Si procederà quindi velocemente ad una sintesi dei concetti già visti. Consideriamo il sistema in figura, che alla serie temporale in ingresso $\{x_n\}$ fa corrispondere, in uscita, la serie temporale $\{y_n\}$.



$$y_n = Q[\{x_n\}]$$

Il generico termine y_n dipende in generale dall'intera serie d'ingresso. Nel caso in cui dipenda solo dai valori passati e presente il sistema si dice causale, in quanto esiste una relazione causa effetto fra ingresso ed uscita, come sempre accade nei sistemi fisicamente realizzabili. Altrimenti il sistema è anticipativo (possono essere anticipativi alcuni sistemi ideali). Se la dipendenza si riduce al solo valore x_n presente in ingresso, il sistema si dice privo di memoria.

Il sistema è detto lineare se:

$$Q[\{c_1 x_n^1 + c_2 x_n^2\}] = c_1 Q[\{x_n^1\}] + c_2 Q[\{x_n^2\}]$$

Il sistema è detto tempo-invariante se la risposta alla serie ritardata è la risposta ritardata, qualunque sia il ritardo (iT):

$$y_{n-i} = Q[\{x_{n-i}\}]$$

Risposta impulsiva

La risposta impulsiva discreta $\{h_n\}$ di un sistema discreto lineare tempo-invariante è la risposta alla serie temporale $\{\delta_n\}$, formata da tutti "0" tranne un "1" nell'origine (corrispondente tempo discreto del Delta di Dirac).

La risposta impulsiva $\{h_n\}$ descrive completamente il comportamento della rete lineare. In particolare consente di esprimere la risposta ad una qualsiasi serie temporale d'ingresso rimanendo nel dominio dei tempi. Infatti, osservando che

$$x_n = \sum_{i=-\infty}^{\infty} x_i \delta_{n-i} = \{x_n\} * \{\delta_n\}$$

Dalle condizioni di linearità e tempo-invarianza, in modo analogo a quanto visto per i sistemi tempo continui, discende la relazione:

$$y_n = \sum_{i=-\infty}^{\infty} x_i h_{n-i} = \{x_n\} * \{h_n\}$$

Funzione di trasferimento

Si definisce funzione di trasferimento $H_s(\omega)$ di un sistema discreto lineare tempo invariante la trasformata secondo Fourier della risposta impulsiva discreta:

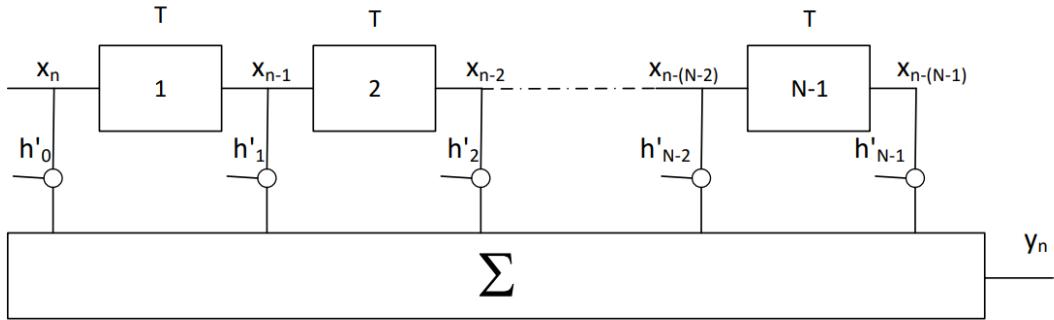
$$H_s(\omega) = F[\{h_n\}] = \sum_{n=-\infty}^{\infty} h_n e^{-in\omega T}$$

Se le serie temporali di ingresso e di uscita sono dotate di trasformata secondo Fourier, per il teorema della convoluzione vale la relazione:

$$Y_s(\omega) = X_s(\omega) H_s(\omega)$$

Filtri trasversali tempo-discreti

Un filtro trasversale tempo-discreto è costituito da un registro a scorrimento (shift-register) e da un sommatore. Di fatto lo schema è lo stesso del caso tempo continuo, tuttavia in ingresso ed in uscita si hanno in questo caso delle serie temporali.



La relazione ingresso-uscita è data da:

$$y_k = \sum_{k=0}^{N-1} h'_k x_{n-k}$$

La risposta impulsiva si ottiene considerando in ingresso la serie $\{\delta_n\}$, di fatto un solo "1" preceduto e seguito da "0". E' immediato verificare che si ottiene una serie data da infiniti "0", seguiti dagli N coefficienti delle prese a partire dall'origine dei tempi, e quindi ancora da infiniti "0":

$$\{h_n\} = \begin{cases} 0 & n < 0 \\ h'_n & 0 \leq n \leq N - 1 \\ 0 & n > N - 1 \end{cases}$$

La funzione di trasferimento è data da

$$H_s(\omega) = F[\{h_n\}] = \sum_{k=0}^{N-1} h'_k e^{-ik\omega T}$$

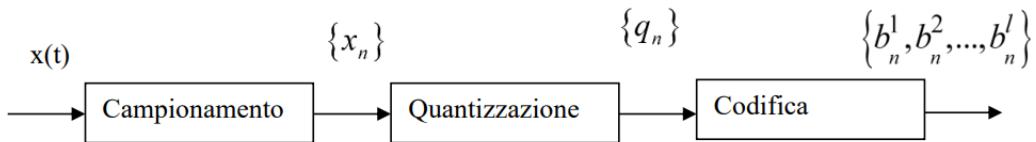
▼ 6.0 - Rappresentazione digitale dei segnali

▼ 6.1 - Conversione analogico digitale

La conversione di un segnale analogico in un segnale digitale (A/D) è di importanza fondamentale in molti dei sistemi moderni. I vantaggi dei segnali digitali sono molteplici, ma forse il più importante è la molto maggiore facilità di elaborazione e di memorizzazione, attuabile con tecniche di tipo informatico.

Le tecniche di conversione analogico digitale sono diverse; qui esamineremo quella di base, detta anche PCM (Pulse Code Modulation) per motivi storici. Da un punto di vista logico la conversione analogico-digitale

PCM prevede tre passaggi: campionamento, quantizzazione e codifica, riassunti nello schema a blocchi della figura sotto.



Campionamento

Per prima cosa è necessario passare da un segnale tempo continuo ad un segnale tempo discreto. Questo è il compito del campionatore, che campionerà quindi il segnale $x(t)$ ad una frequenza di campionamento f_o , ottenendo la serie temporale $\{x_n\}$ $x_n = x(nT)$.

Il teorema di Shannon ci assicura che non si ha perdita di informazione, cioè l'operazione è reversibile, se $f_o > 2f_m$. Il rispetto della condizione può avvenire in due modi: o aumentando la frequenza di campionamento fino a soddisfare la condizione (con un certo margine, necessario per l'operazione inversa), oppure, nel caso in cui la frequenza di campionamento sia fissa, anteponendo al campionatore un filtro passa-basso in grado di ridurre la frequenza massima del segnale.

Quantizzazione

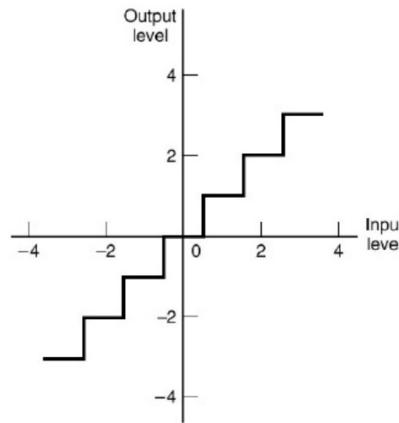
La serie temporale $\{x_n\}$ è un segnale tempo-discreto, ma non discreto nei valori. Assumendo $x(t)$ bilanciato, con valori compresi nell'intervallo $[-M, M]$, i campioni risulteranno anch'essi compresi in detto intervallo, potendo assumere qualsiasi valore all'interno di esso. Per poter procedere, è necessario ridurre il numero dei valori da infinito a finito, passando da una serie temporale continua nei valori ad una discreta nei valori. L'operazione viene detta quantizzazione, e consiste in una approssimazione dei valori ottenuti. L'intervallo di variabilità $[-M, M]$ dei valori campionati viene suddiviso in un numero finito di intervalli (intervalli di quantizzazione) e tutti i valori interni a ciascuno di questi vengono identificati con uno di essi, che indichiamo con q_n . L'operazione di quantizzazione non è evidentemente reversibile, in quanto non è possibile risalire da un valore quantizzato al campione che l'ha generato.

L'intervallo su cui opera il quantizzatore, $[-M_q, M_q]$ viene detto dinamica del quantizzatore, ed è opportuno che coincida con quella del segnale, come verrà mostrato in seguito. Spesso, ma non necessariamente, la dinamica del quantizzatore è suddivisa in intervalli uguali, nel qual caso si dice "uniforme". In caso contrario, si dice "non-uniforme". Il valore

rappresentativo di ogni intervallo è arbitrario, ma spesso viene scelto per simmetria il valore centrale. La differenza fra campione e valore quantizzato corrispondente si dice "errore" di quantizzazione:

$$e_n = x_n - q_n$$

L'operazione di quantizzazione può essere rappresentata nel seguente modo:



Il legame funzionale fra campioni e valori quantizzati è detta legge di quantizzazione:

$$q_n = f(x_n)$$

I valori assumibili dalla variabile q_n vengono anche detti "livelli" di quantizzazione.

La scelta del quantizzatore ottimo, cioè della miglior legge di quantizzazione, intendendo per migliore quella che minimizza gli effetti dell'errore di quantizzazione, dipende dalla statistica dei campioni. Per la sua semplicità, viene spesso usato il quantizzatore uniforme, anche quando non ottimo. Risulta evidente che maggiore è il numero di livelli, L , minore sarà in generale l'errore di quantizzazione. L'aumento di L , tuttavia, comporta come vedremo fra un attimo un aumento dei bit necessari a rappresentare il segnale, per cui si impone un compromesso fra qualità e numero di bit.

Codifica

I primi due passi della conversione A/D ci permettono di ottenere una serie tempo discreta e discreta nei valori, a partire da un segnale analogico, quindi tempo continuo e continuo nei valori. Il passo successivo è trasformare la serie dei valori quantizzati in una serie di bit. La codifica

associa ad ognuno degli L livelli che possono essere assunti dai valori quantizzati una parola formata da un certo numero di bit, in modo da avere una corrispondenza biunivoca fra valori ed “etichette” binarie. Non considerando qui volutamente le codifiche entropiche per semplicità, si assume che tutte le etichette siano formate dallo stesso numero di bit l , che dovrà essere quindi messo in relazione al numero di livelli L :

$$l \geq \log_2 L$$

Di norma conviene prendere L potenza di due o potenza di due meno uno, in modo da avere rispettivamente $L = 2^l$ o $L = 2^l - 1$. La prima scelta è preferita quando si vuole che l’origine si trovi al confine fra due intervalli diversi (“quantizzatore “midrise”), la seconda quando si vuole inserire un intervallo a cavallo dell’origine (quantizzatori “midtread”). La seconda scelta ha il vantaggio di rendere nullo il segnale ricostruito quando i campioni sono molto piccoli e cioè vicinissimi all’origine (ad esempio per eliminare il fruscio in assenza di segnale utile).

Ciò premesso, il codificatore assocerà ad ogni elemento della serie $\{q_n\}$ una parola di n bit, ottenendo una serie di parole binarie $\{b_n^1 b_n^2 \dots b_n^l\}$. Conoscendo il periodo di campionamento T è possibile calcolare il tempo medio necessario per codificare ciascun bit del segnale analogico in un valore digitale, ovvero il tempo di bit:

$$T_b = T/l$$

e la conseguente frequenza di bit:

$$f_b = 1/T_b \implies f_b = f_o l$$

Essa rappresenta il numero di bit necessari a rappresentare (quindi sia a trasmettere che memorizzare) un secondo del segnale, ed è evidentemente un fattore fondamentale nella valutazione di un sistema di conversione A/D.

▼ 6.2 - Conversione digitale analogico

La conversione digitale analogica (D/A) consiste nella ricostruzione del segnale originario $x(t)$ a partire dal corrispondente messaggio numerico, cioè dai bit che lo rappresentano. Essa prevede due soli passi: decodifica e ricostruzione del segnale.

Decodifica

È ovviamente il processo inverso della codifica, mediante il quale si ricostruiscono i valori quantizzati $\{q_n\}$, a partire dalle parole $\{b_n^1 b_n^2 \dots b_n^l\}$. Essendo la quantizzazione una operazione non reversibile, la serie $\{q_n\}$ deve necessariamente essere trattata come se fosse la serie $\{x_n\}$ dei valori compionati. L'errore di quantizzazione farà però sì che il segnale ricostruito differisca da quello originale: la differenza prenderà il nome di rumore di quantizzazione:

$$e(t) = x_r(t) - x(t)$$

Ricostruzione del segnale

Generazione del segnale PAM

Dal punto di vista teorico il segnale potrebbe essere ricostruito a partire dalla conoscenza dei suoi campioni utilizzando la serie di Shannon. Risulta tuttavia molto più conveniente ricostruire il segnale a partire dal segnale PAM ottenuto come prodotto di convoluzione della serie dei valori quantizzati, equivalenti ai campioni, a meno dell'errore dovuto alla quantizzazione (qui momentaneamente trascurato per poter procedere con la dimostrazione), con un impulso rettangolare $g(t)$, di ampiezza unitaria, con origine a $t=0$. Si ottiene quindi il segnale PAM:

$$s(t) = \{x_n\} * g(t) = \sum_{-\infty}^{\infty} x_n g(t - nT)$$

Filtratura passa-basso

Mostriamo ora come sia possibile ricostruire il segnale $x(t)$ a partire da quello PAM. Richiamando i risultati ottenuti per la trasformata di un segnale PAM si ha:

$$S(\omega) = X_s(\omega)G(\omega) = \frac{1}{T} \sum_{k=-\infty}^{\infty} X(\omega + k\omega_o)G(\omega)$$

Se il campionamento di $x(t)$ è stato effettuato in accordo con la condizione del teorema di campionamento, i termini della ripetizione periodica della trasformata del segnale non si sovrappongono (assenza di aliasing nelle frequenze) e di conseguenza l'impiego di un filtro passabasso consente di isolare il termine centrale. La frequenza di campionamento è bene sia un po' superiore al doppio della frequenza del segnale per consentire un raccordo continuo da banda passante $([0, f_m])$ a banda attenuata $[f_o - f_m, \infty]$ (si

ricorda che un filtro reale non può avere caratteristica di ampiezza discontinua). Si ha quindi in uscita al filtro (trascurando il ritardo da esso introdotto):

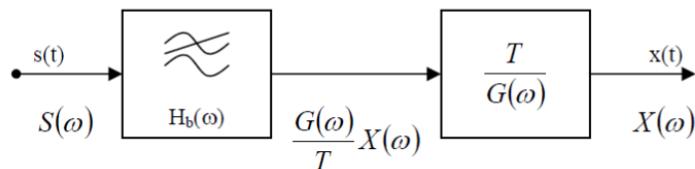
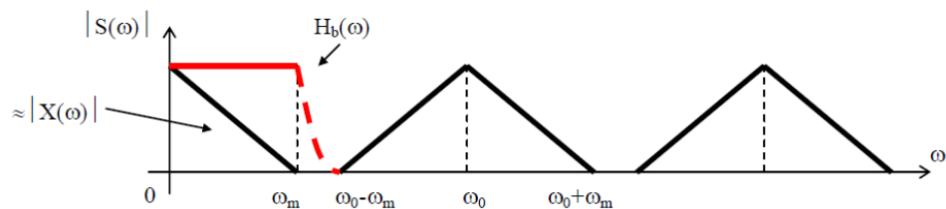
$$S_f(\omega) = X(\omega) \frac{G(\omega)}{T}$$

Eventuale equalizzazione

Per ottenere il segnale originario occorre porre in cascata al filtro passa-basso una rete equalizzatrice la cui funzione di trasferimento è data da:

$$H_e(\omega) = \begin{cases} \frac{T}{G(\omega)} & |\omega| \leq \omega_m \\ \text{qualsiasi} & |\omega| \leq \omega_m \end{cases}$$

Tuttavia, se la durata τ dell'impulso $g(t)$ è molto minore di T , la sua trasformata $G(\omega)$ è praticamente costante ed eguale a τ nella banda del segnale, in quanto ha un andamento a lobi e il primo punto di nullo è a $2\pi/\tau \gg 2\pi/T = \omega_o > 2\omega_m$. La rete equalizzatrice può quindi essere omessa, dato che si ottiene comunque $x(t)$ in uscita al filtro passa-basso a meno di una costante moltiplicativa. In altri termini, un filtro passa-basso è sufficiente per ricostruire dal segnale PAM il segnale originario a meno di una costante moltiplicativa e del ritardo introdotto dal circuito.



▼ 7.0 - Introduzione alla teoria della modulazione

▼ 7.1 - Modulazioni

In alcuni casi, ed in particolare in quello radio, può capitare che si debba trasmettere un segnale su un canale di tipo passa-banda, mentre il segnale è di tipo passa-basso.

In generale, tutte le volte che il segnale $x(t)$ da trasmettere è passa-basso mentre il canale di trasmissione è passa-banda, occorre associare al segnale $x(t)$ un segnale passa-banda, trasmissibile sul canale. Tale associazione deve naturalmente essere invertibile, per restituire all'utilizzatore finale un segnale passa-basso. Questa associazione si ottiene ricorrendo alla modulazione di una oscillazione sinusoidale, a frequenza sufficientemente elevata, detta "portante". Il segnale $x(t)$ si dice allora segnale "modulante", in quanto modula, cioè modifica, le caratteristiche della portante (ampiezza e/o argomento), mentre il segnale $s(t)$ così ottenuto è detto oscillazione "modulata".

Una modulazione si dice analogica o digitale a seconda che sia analogico o digitale (tipo PAM) il segnale modulante. Nel seguito la trattazione si limiterà alle sole modulazioni analogiche per motivi di tempo. Quelle digitali possono essere tuttavia ottenute in linea di principio sostituendo al segnale $x(t)$ analogico un segnale digitale.

Definizioni

La portante, come ogni sinusoide, è caratterizzata dai parametri ampiezza, pulsazione e fase. Dato che nel processo di modulazione uno o più di questi parametri vengono modificati dal segnale modulante, essi vengono detti "iniziali" e contraddistinti dal pedice "o", che contraddistingue la portante stessa. L'espressione della portante è quindi:

$$s_o(t) = V_0 \cos(\omega_o t - \varphi_o)$$

Per trasmettere informazione bisogna variare (modulare) uno o più dei parametri sopra menzionati. Tali variazioni debbono inoltre essere "lente" rispetto alla rapidità di variazione della portante; a tal fine basta scegliere la pulsazione della portante stessa sufficientemente elevata rispetto alla massima pulsazione contenuta nello spettro del segnale modulante. Si ottiene in tal modo un'oscillazione sinusoidale modulata $s(t)$ la cui espressione più generale può essere posta nella forma:

$$s(t) = V(t) \cos \varphi(t) \quad V(t) \geq 0$$

Diamo ora alcune definizioni:

- $V(t) - V_o$: deviazione istantanea di ampiezza
- $m(t) = \frac{V(t) - V_o}{V_o}$: deviazione istantanea relativa di ampiezza (poichè $V(t) \geq 0$, segue $m(t) \geq -1$)

- $\alpha(t) = \varphi(t) - (\omega_o t - \varphi_o)$: deviazione istantanea di fase
- $\Delta\omega(t) = \omega(t) - \omega_o$: deviazione istantanea di pulsazione

Le deviazioni di fase e di pulsazione risultano legate:

$$\begin{aligned}\Delta\omega(t) &= \dot{\alpha}(t) \\ \alpha(t) &= \int_{-\infty}^t \Delta\omega(\tau) d\tau\end{aligned}$$

AM: modulazione di ampiezza

$$AM : \begin{cases} m(t) = kx(t) \\ \alpha(t) = 0 \end{cases}$$

La deviazione relativa di ampiezza è proporzionale al segnale modulante; la deviazione di fase è nulla. Viene modificata quindi solo l'ampiezza dell'oscillazione portante.

$$s(t) = V_o[1 + kx(t)] \cos(\omega_o t - \varphi_o)$$

PM: modulazione di fase

$$PM : \begin{cases} m(t) = 0 \\ \alpha(t) = kx(t) \end{cases}$$

La deviazione relativa di ampiezza è nulla, la deviazione di fase è proporzionale al segnale modulante.

$$s(t) = V_o \cos[\omega_o t + kx(t) - \varphi_o]$$

FM: modulazione di frequenza

$$FM : \begin{cases} m(t) = 0 \\ \Delta\omega(t) = kx(t) \end{cases}$$

E' simile alla PM, ma in questo caso è la deviazione istantanea di pulsazione ad essere direttamente proporzionale al segnale modulante e non quella di fase.

$$s(t) = V_o \cos[\omega_o t + k \int_{-\infty}^t x(\tau) d\tau - \varphi_o]$$

Indice di modulazione

L'indice di modulazione indica il livello di modulazione (cioè l'entità della trasformazione) che viene impressa alla portante dal segnale modulante. L'indice assume il valore 0 in assenza totale di modulazione, ed il valore 1 quando si ha il massimo della modulazione.

Indice di modulazione di ampiezza

In AM l'indice di modulazione è definito come

$$m_a = \max(|m(t)|)$$

Occorre quindi prestare attenzione a non portare il modulatore in "sovramodulazione", cioè a non avere $kM > 1$ (o $m_a > 1$). Nel caso in cui si superi tale valore il segnale va in "sovramodulazione" e la modulazione diventa ibrida. Infatti in questo caso si ha:

$$\text{sovramodulazione AM : } \begin{cases} V(t) = V_o |1 + kx(t)| \\ \alpha(t) = \begin{cases} 0 & 1 + kx(t) > 0 \\ \pi & 1 + kx(t) < 0 \end{cases} \end{cases}$$

Indice di modulazione d'angolo

Nelle modulazioni d'angolo l'indice di modulazione è definito come

$$m = \max(|\alpha(t)|)$$

Si noti che non esiste alcuna limitazione superiore a differenza dell'AM.

Inviluppo complesso rappresentativo (o equivalente passa-basso) di oscillazioni modulate

Consideriamo una nuova espressione generale di un'oscillazione sinusoidale modulata

$$s(t) = V(t) \cos[\omega_o t + \alpha(t) - \varphi_o]$$

Essa, in analogia con quanto fatto nel metodo simbolico di Steinmetz, può essere scritta nella forma:

$$s(t) = \operatorname{Re}\{i(t)e^{i\omega_o t}\} \quad i(t) = V(t)e^{i[\alpha(t)-\varphi_o]}$$

Nota la frequenza della portante, l'oscillazione è completamente individuata dal suo inviluppo complesso rappresentativo $i(t)$. $i(t)$ è passa-basso anziché

passa-banda. Per questo motivo viene anche chiamato "equivalente passa-basso" di $s(t)$. Si noti infine che se non si ha nessuna modulazione, allora l'inviluppo complesso da funzione diventa una costante, ricadendo nel caso del metodo simbolico classico.

Consideriamo ora due oscillazioni modulate diverse, ma aventi la stessa pulsazione della portante

$$\begin{aligned} s_1(t) &= \operatorname{Re}\{i_1(t)e^{i\omega_o t}\} & i_1(t) &= V_1(t)e^{i[\alpha_1(t)-\varphi_{o1}]} \\ s_2(t) &= \operatorname{Re}\{i_2(t)e^{i\omega_o t}\} & i_2(t) &= V_2(t)e^{i[\alpha_2(t)-\varphi_{o2}]} \end{aligned}$$

Per la somma risulta

$$\begin{aligned} s(t) &= s_1(t) + s_2(t) = \operatorname{Re}\{i_1(t)e^{i\omega_o t}\} + \operatorname{Re}\{i_2(t)e^{i\omega_o t}\} \\ &= \operatorname{Re}\{[i_1(t) + i_2(t)]e^{i\omega_o t}\} = \operatorname{Re}\{i(t)e^{i\omega_o t}\} \end{aligned}$$

Dove l'inviluppo complesso della somma è dato dalla somma dei due inviluppi,

$$i(t) = i_1(t) + i_2(t)$$

Dalle due formule precedenti si vede che il segnale $s(t)$ è anch'esso una oscillazione modulata e che il suo inviluppo complesso è uguale alla somma degli inviluppi complessi delle oscillazioni modulate componenti. Da $i(t)$ è facile ricavare l'ampiezza istantanea e la fase istantanea della modulazione $s(t)$, cioè le "leggi" della modulazione. Si ha infatti:

$$\begin{cases} V(t) = |i(t)| \\ \alpha(t) = \arg\{i(t)\} \end{cases}$$

La formula può naturalmente essere estesa a più segnali, eventualmente moltiplicati per delle costanti. Può anche essere utilizzata in senso inverso, per scomporre un'oscillazione modulata in più oscillazioni modulate componenti. L'equivalente passa-basso è praticamente sempre utilizzato al posto di $s(t)$ nello studio delle oscillazioni modulate, sia analogiche che numeriche.

▼ 7.2 - Caratteristiche di un'oscillazione

Caratteristiche spettrali di una oscillazione AM. Oscillazioni DSB, SSB, DSB-SC, SSB-SC.

DSB

Allo scopo di determinare lo spettro di una modulazione AM consideriamo di nuovo la sua espressione, assunta per semplicità nulla la fase della portante:

$$s(t) = V_o[1 + kx(t)] \cos \omega_o t$$

Separando i termini di $s(t)$ possiamo vedere immediatamente che è dato dalla somma della portante e di un termine che, a meno di una costante, è il prodotto fra il segnale modulante e la portante.

$$s(t) = V_o \cos \omega_o t + V_o kx(t) \cos \omega_o t$$

Supponiamo che il segnale modulante $x(t)$, di tipo passa-basso, sia rappresentabile mediante l'integrale di Fourier:

$$x(t) = \int_{\omega_i}^{\omega_m} V(\omega) \cos[\omega t - \varphi(\omega)] d\omega$$

Sostituendo la formula sopra nell'espressione generale dell'AM, è possibile scomporre l'AM nella somma di tre termini: portante, banda laterale superiore, banda laterale inferiore:

$$\begin{aligned} s(t) = & V_o \cos \omega_o t + \frac{kV_o}{2} \int_{\omega_i}^{\omega_m} V(\omega) \cos[(\omega_o + \omega)t - \varphi(\omega)] d\omega + \\ & + \frac{kV_o}{2} \int_{\omega_i}^{\omega_m} V(\omega) \cos[(\omega_o - \omega)t + \varphi(\omega)] d\omega \end{aligned}$$

A titolo di esempio è riportato nella figura sotto un possibile andamento degli spettri di ampiezza e di fase di una generica modulazione AM.

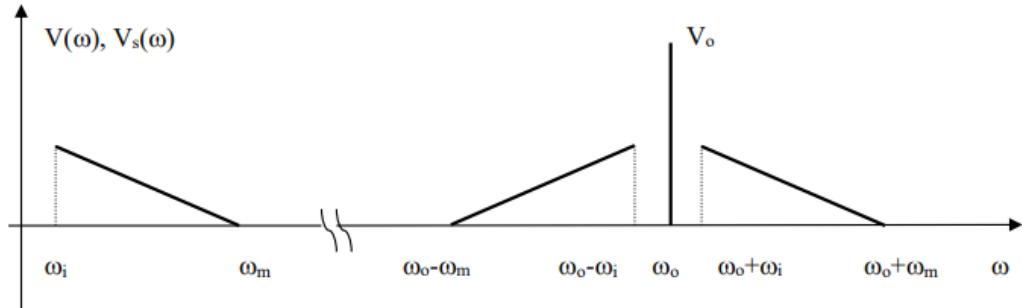


Fig.106 Spettro di ampiezza AM.

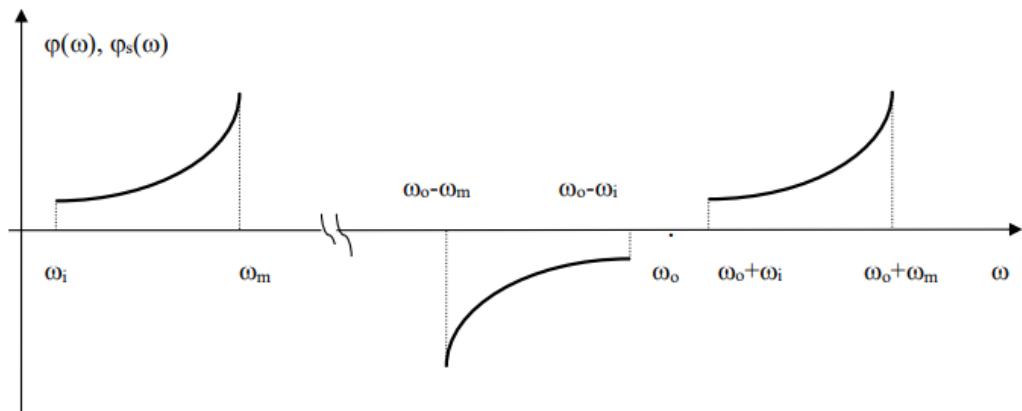


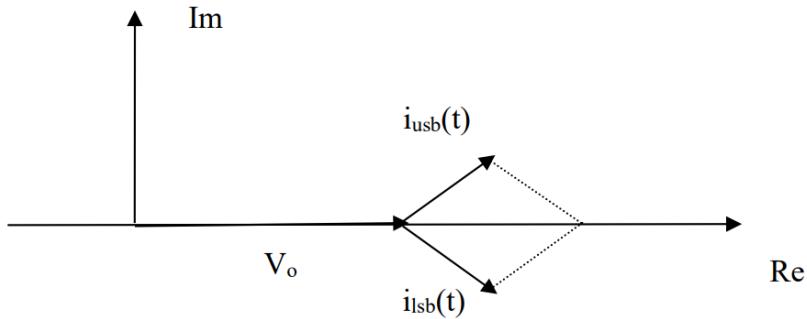
Fig.107 Spettro di fase AM

L'oscillazione modulata occupa una banda di pulsazioni centrata sulla pulsazione della portante, di larghezza $B_\omega = 2\omega_m$. Lo spettro di ampiezza è simmetrico rispetto alla pulsazione della portante, mentre quello di fase è antisimmetrico. Nell'ipotesi, normalmente verificata, che sia $\omega_m \ll \omega_o$ risulta $B_\omega \ll \omega_o$, cioè l'oscillazione AM è un segnale passa-banda.

Dall'espressione precedente è possibile ricavare l'inviluppo complesso dell'AM come somma degli inviluppi complessi delle tre componenti:

$$i(t) = V_o + \frac{kV_o}{2} \int_{\omega_i}^{\omega_m} V(\omega) e^{j[\omega t - \varphi(\omega)]} d\omega + \frac{kV_o}{2} \int_{\omega_i}^{\omega_m} V(\omega) e^{-j[\omega t - \varphi(\omega)]} d\omega$$

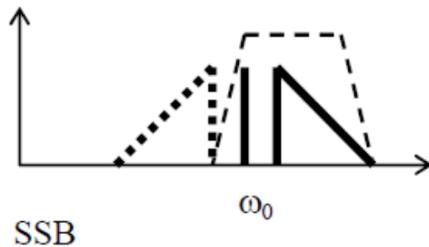
Si noti che il secondo ed il terzo termine ($i_{usb}(t)$, e $i_{lsb}(t)$) sono complessi coniugati, per cui la loro risultante giace sull'asse reale; essa può essere positiva o negativa e si somma a V_o . La risultante totale non può mai essere negativa, altrimenti avremmo una modulazione ibrida.



SSB

Poiché ciascuno dei due segnali corrispondenti alla banda laterale superiore od inferiore contiene la stessa informazione (nota una delle due bande laterali, è immediato ricavare l'altra), è possibile eliminare una delle due bande laterali con l'evidente vantaggio di dimezzare la banda impegnata dal relativo segnale modulato.

Il segnale modulato che così si ottiene viene chiamato a banda laterale singola ed indicato con la sigla SSB (Single Side Band). Esso può essere generato a partire da una oscillazione AM, che viene filtrata con un filtro passa-banda che provveda ad eliminare la banda indesiderata.



Le espressioni della SSB sono:

$$s(t) = V_o \cos \omega_o t + \frac{kV_o}{2} \int_{\omega_l}^{\omega_m} V(\omega) \cos[(\omega_o + \omega)t - \varphi(\omega)] d\omega$$

$$i(t) = V_o + \frac{kV_o}{2} \int_{\omega_l}^{\omega_m} V(\omega) e^{j[\omega t - \varphi(\omega)]} d\omega$$

Si noti dall'espressione dell'inviluppo complesso che l'eliminazione di una delle due bande laterali rende la modulazione SSB ibrida, in quanto varia sia il modulo che l'argomento di $i(t)$.

DSB-SC

Una seconda modulazione derivata dall'AM è la DSB-SC (DSB Suppressed Carrier). In essa vengono mantenute entrambe le bande laterali, ma viene eliminata la portante, per risparmiare potenza, senza perdere il contenuto informativo relativo al segnale modulante. Le espressioni della DSB-SC sono:

$$s(t) = \frac{kV_o}{2} \int_{\omega_l}^{\omega_m} V(\omega) \cos[(\omega_o + \omega)t - \varphi(\omega)] d\omega + \frac{kV_o}{2} \int_{\omega_l}^{\omega_m} V(\omega) \cos[(\omega_o - \omega)t + \varphi(\omega)] d\omega$$

$$i(t) = \frac{kV_o}{2} \int_{\omega_l}^{\omega_m} V(\omega) e^{j[\omega t - \varphi(\omega)]} d\omega + \frac{kV_o}{2} \int_{\omega_l}^{\omega_m} V(\omega) e^{-j[\omega t - \varphi(\omega)]} d\omega = kV_o x(t)$$

L'eliminazione della portante rende la DSB-SC ibrida. Infatti, pur rimanendo $i(t)$ reale (sotto l'ipotesi di fase della portante nulla), esso può essere positivo o negativo, rendendo non identicamente nulla la deviazione istantanea di fase.

SSB-SC (conversione di frequenza)

Se a partire da una DSB-SC si elimina anche una delle due bande laterali mediante un filtro, si ottiene una modulazione SSB-SC (SSB - Suppressed Carrier). In questo caso il vantaggio è il dimezzamento della banda occupata, da sommare al risparmio in potenza già ottenuto nella DSB eliminando la portante. Le espressioni della SSB-SC sono:

$$s(t) = \frac{kV_o}{2} \int_{\omega_l}^{\omega_m} V(\omega) \cos[(\omega_o + \omega)t - \varphi(\omega)] d\omega$$

$$i(t) = \frac{kV_o}{2} \int_{\omega_l}^{\omega_m} V(\omega) e^{j[\omega t - \varphi(\omega)]} d\omega$$

La SSB-SC equivale ad una conversione di frequenza in salita, in quanto gli spettri di ampiezza e fase dell'oscillazione modulata coincidono con quelli del segnale modulante, traslati di ω_o .

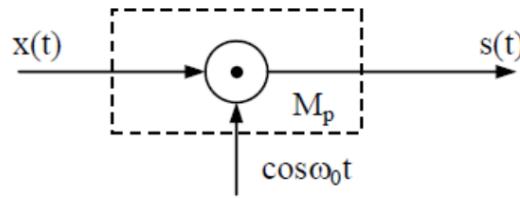
La SSB-SC e la sua variante con soppressione parziale sono modulazioni ibride.

Modulazioni a prodotto e QAM

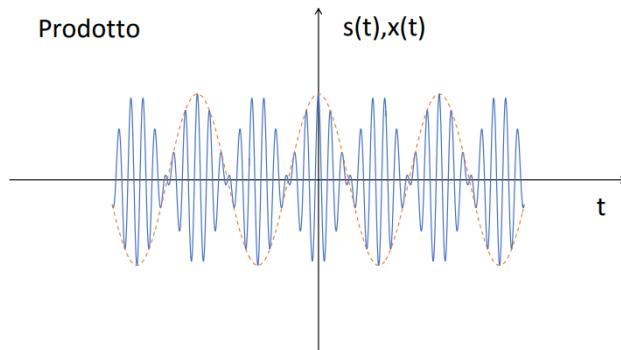
Modulazioni a prodotto

La modulazione a prodotto può essere ottenuta direttamente, senza passare dall'AM, facendo il prodotto del segnale modulante $x(t)$ per una sinusoida tramite un circuito detto modulatore a prodotto o "mixer", ottenendo:

$$s(t) = x(t) \cos \omega_o t$$



Un possibile andamento dell'oscillazione a prodotto è il seguente:



L'inviluppo complesso è estremamente semplice in quanto coincide con il segnale modulante:

$$i(t) = x(t)$$

La modulazione è ibrida. Infatti sia l'ampiezza istantanea che la deviazione istantanea di fase variano nel tempo:

$$V(t) = |i(t)|$$

$$\alpha(t) = \begin{cases} 0 & x(t) > 0 \\ \pi & x(t) < 0 \end{cases}$$

Gli spettri si ottengono immediatamente dall'espressione della trasformata (teorema fondamentale della modulazione), ricordando che essendo in questo caso $\omega_m \ll \omega_o$ i due termini non sono sovrapposti ed il segnale è passa-banda.

$$S(\omega) = \frac{1}{2}X(\omega - \omega_o) + \frac{1}{2}X(\omega + \omega_o)$$

La banda di $s(t)$ è doppia rispetto a quella di $x(t)$.

Demodulazioni a prodotto

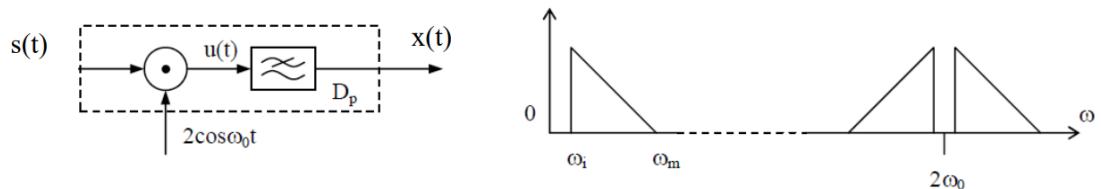
La demodulazione si ottiene rimoltiplicando l'oscillazione modulata per la portante (moltiplicata per due per convenienza formale)

$$u(t) = 2s(t) \cos \omega_o t = 2x(t) \cos^2 \omega_o t = x(t) + x(t) \cos 2\omega_o t$$

e filtrando passa-basso nella banda di $x(t)$ il segnale ottenuto per eliminare il secondo termine.

$$x_d(t) = x(t)$$

Il demodulatore è quindi costituito da un modulatore a prodotto seguito da un filtro passa-basso.



L'errore di fase provoca un'attenuazione del segnale demodulato, che sarà:

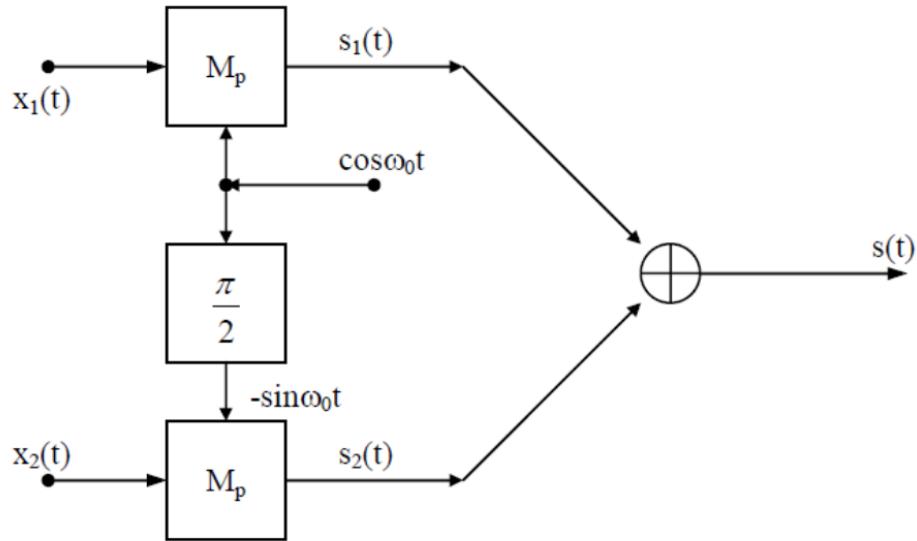
$$x_d(t) = x(t) \cos \Delta$$

Modulazioni QAM (Quadrature Amplitude Modulation)

La modulazione QAM è un'estensione della modulazione a prodotto. Essa infatti consiste di due modulazioni a prodotto con portanti in quadratura, cioè con la seconda portante sfasata in anticipo di $\pi/2$. Il segnale in uscita al primo modulatore si chiama via in fase, l'altro via in quadratura.

Vi sono due segnali modulanti $x_1(t)$ e $x_2(t)$, aventi le stesse caratteristiche spettrali; devono inoltre essere "indipendenti", cioè non legati fra di loro da nessuna relazione. Si ha:

$$s(t) = x_1(t) \cos \omega_o t - x_2(t) \sin \omega_o t$$



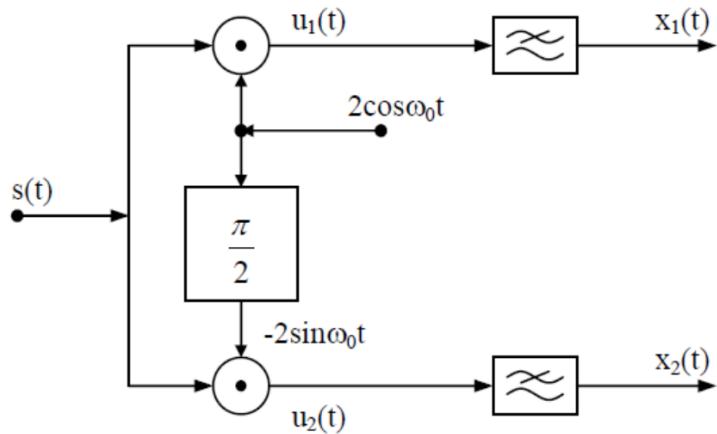
L'inviluppo complesso è dato da:

$$i(t) = x_1(t) + jx_2(t)$$

La modulazione è ibrida, infatti $i(t)$ varia sia in modulo che in argomento.

Gli spettri della via in fase e della via in quadratura si sovrappongono. E' quindi possibile trasmettere il doppio di informazione nella stessa banda di una modulazione a prodotto.

Il demodulatore è la somma di due demodulatori a prodotto.



Per il segnale in uscita al primo (componente in fase) si ha:

$$\begin{aligned} u_p(t) &= 2s(t)\cos \omega_o t = 2x_1(t)\cos^2 \omega_o t - 2x_2(t)\sin \omega_o t \cos \omega_o t = \\ &= x_1(t) + x_1(t)\cos 2\omega_o t - x_2(t)\sin 2\omega_o t \end{aligned}$$

La prima delle tre componenti è l'unica voluta e viene isolata dal filtro passabasso.

$$x_{pd}(t) = x_1(t)$$

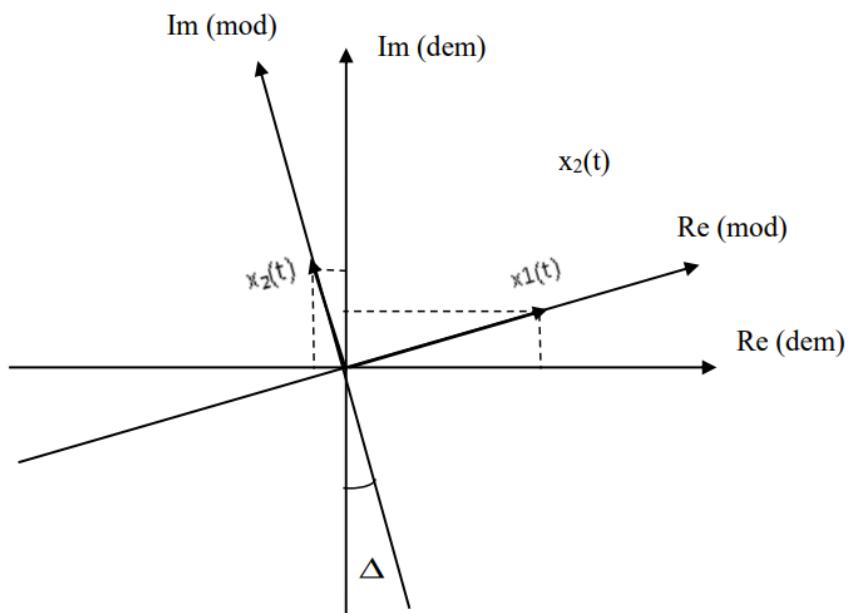
Analogamente per la via in quadratura si ha:

$$\begin{aligned} u_q(t) &= -2s(t)\sin\omega_o t = 2x_2(t)\sin^2\omega_o t - 2x_1(t)\sin\omega_o t \cos\omega_o t = \\ &= x_2(t) - x_2(t)\cos 2\omega_o t - x_1(t)\sin 2\omega_o t \end{aligned}$$

$$x_{qd}(t) = x_2(t)$$

Nel caso del QAM un errore di fase provoca non solo un'attenuazione del segnale utile, ma anche un'interferenza dalla via in quadratura. Infatti nel caso di una sfasatura in ritardo si ha:

$$\begin{aligned} x_{pd}(t) &= x_1(t)\cos\Delta - x_2(t)\sin\Delta \\ x_{qd}(t) &= x_2(t)\cos\Delta + x_1(t)\sin\Delta \end{aligned}$$



Spettro segnali modulati in angolo: formula di Carson

Per calcolare lo spettro delle oscillazioni modulate in angolo (PM, FM) è possibile utilizzare la formula di Carson. Sia dunque ω_m la massima pulsazione del segnale modulante e $\Delta\omega_{max} = \max(\Delta\omega(t))$ la massima deviazione di frequenza, la banda risulta approssimativamente data da

$$B_\omega = 2(\omega_m + \Delta\omega_{max})$$

Nei casi pratici la formula di Carson vien data nelle frequenze, con ovvio significato dei simboli.

$$B = 2(f_m + \Delta f_{max})$$

▼ 8.0 - Segnali ad energia ed a potenza finita

Energia e potenza di un segnale

Dato un segnale $x(t)$, in generale complesso, si definisce "potenza istantanea"

$$p(t) = x^*(t)x(t) = |x(t)|^2$$

da cui derivano le seguenti definizioni di "energia" e "potenza" (media)

$$E = \int_{-\infty}^{+\infty} |x(t)|^2 dt$$

$$P = \langle |x(t)|^2 \rangle = \lim_{T \rightarrow \infty} \frac{1}{T} \int_{-\frac{T}{2}}^{+\frac{T}{2}} |x(t)|^2 dt$$

Un segnale si dice ad energia finita se l'integrale che ne rappresenta l'energia converge, mentre è a potenza finita nel caso in cui converga, ad un valore diverso da zero, l'integrale che ne rappresenta la potenza. Si noti che per un segnale ad energia finita la potenza tende a zero, per cui un segnale non può appartenere ad entrambe le categorie. Inoltre esistono segnali che non sono né ad energia né a potenza finita.

Valore efficace

Per un segnale a potenza finita si definisce valore efficace la costante che ha la stessa potenza del segnale, ovvero

$$x_{eff} = \sqrt{P}$$

Analisi generalizzata

I segnali ad energia finita ammettono trasformata di Fourier, mentre quelli a potenza finita in generale no, ad eccezione dei segnali periodici.

Nei prossimi paragrafi verrà introdotta tramite gli spettri di energia e di potenza una rappresentazione "energetica" dei segnali nel dominio delle frequenze.

Questa rappresentazione si aggiunge per i segnali ad energia finita alla rappresentazione usuale mediante trasformata di Fourier, mentre spesso è

l'unica possibile per i secondi. Si parla quindi di analisi di Fourier generalizzata, per indicare che la possibilità di una rappresentazione nel dominio delle frequenze viene estesa anche a molti segnali che non ammettono trasformata di Fourier.

Segnali ad energia finita

Funzioni di crosscorrelazione ed autocorrelazione

Dati due segnali in generale complessi, $x(t)$ ed $y(t)$, si definisce funzione di crosscorrelazione il coniugato del prodotto interno di un uno di essi per la versione anticipata dell'altro:

$$\dot{\varphi}_{xy}(\tau) = \langle x, y_\tau \rangle^* = \int_{-\infty}^{+\infty} x^*(t)y(t + \tau) dt$$

Nel caso particolare in cui $y(t)=x(t)$, la funzione di crosscorrelazione prende il nome di funzione di autocorrelazione $\dot{\varphi}_x(\tau)$. Si noti che l'autocorrelazione calcolata nell'origine rappresenta l'energia di un segnale:

$$\dot{\varphi}_x(0) = \int_{-\infty}^{+\infty} x^*(t)x(t) dt = \int_{-\infty}^{+\infty} |x(t)|^2 dt = E_x$$

Proprietà delle funzioni di cross e autocorrelazione

Per la funzione di crosscorrelazione e autocorrelazione vale la seguente proprietà:

$$\begin{aligned}\dot{\varphi}_{xy}(\tau) &= \dot{\varphi}_{yx}^*(-\tau) \\ \dot{\varphi}_x(\tau) &= \dot{\varphi}_x^*(-\tau)\end{aligned}$$

Dimostrazione pag. 123

Applicando la diseguaglianza di Schwarz alla definizione di crosscorrelazione si ha:

$$\begin{aligned}|\dot{\varphi}_{xy}(\tau)|^2 &\leq \dot{\varphi}_x(0)\dot{\varphi}_y(0) = E_x E_y \\ |\dot{\varphi}_x(\tau)| &\leq \dot{\varphi}_x(0) = E_x\end{aligned}$$

La relazione appena trovata stabilisce che il massimo (in modulo) della funzione di autocorrelazione è nell'origine, e che esso coincide con l'energia.

Un'altra importante relazione è la seguente:

$$\begin{aligned}\dot{\varphi}_{xy}(\tau) &= x^*(-\tau) * y(\tau) \\ \dot{\varphi}_x(\tau) &= x^*(-\tau) * x(\tau)\end{aligned}$$

Dimostrazione pag. 124

Teorema di Parseval generalizzato e condizioni di ortogonalità

Il teorema di Parseval generalizzato afferma che:

$$\int_{-\infty}^{+\infty} x^*(t)y(t) dt = \frac{1}{2\pi} \int_{-\infty}^{+\infty} X^*(\omega)Y(\omega) d\omega$$

▼ Dimostrazione

Per le funzioni ad energia finita è garantita l'esistenza della trasformata di Fourier. Applicando l'operatore trasformata di Fourier alla funzione di crosscorrelazione si ottiene

$$\dot{\Phi}_{xy}(\omega) = F[\dot{\varphi}_{xy}(\tau)] = F[x^*(-\tau)]F[y(\tau)] = X^*(\omega)Y(\omega)$$

Ora è possibile scrivere la funzione di crosscorrelazione come antitrasformata della sua trasformata

$$\dot{\varphi}_{xy}(\tau) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} \dot{\Phi}_{xy}(\omega)e^{j\omega\tau} d\omega$$

e valutare il valore assunto nell'origine nel modo seguente

$$\dot{\varphi}_{xy}(0) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} \dot{\Phi}_{xy}(\omega)d\omega = \frac{1}{2\pi} \int_{-\infty}^{+\infty} X^*(\omega)Y(\omega)d\omega$$

ottenendo dunque l'enunciato del teorema di Parseval generalizzato.

QED.

Si noti che il primo termine rappresenta il prodotto interno coniugato di $x(t)$ e $y(t)$. In analogia con il prodotto interno di due vettori, se esso si annulla i segnali $x(t)$ e $y(t)$ si dicono ortogonali. Dal teorema di Parseval generalizzato si deduce che se due segnali sono ortogonali, lo devono per forza essere anche le relative trasformate, e viceversa. Inoltre si possono fare le seguenti osservazioni:

- condizione sufficiente affinché due segnali siano ortogonali è che essi non si sovrappongano nel tempo (dal primo integrale)
- condizione sufficiente affinché due segnali siano ortogonali è che essi non si sovrappongano in frequenza (dal secondo integrale)

- due segnali che si sovrappongono sia nel tempo che in frequenza possono ancora essere ortogonali

Densità spettrale di energia

Dal teorema di Parseval si ottiene che:

$$E_x = \int_{-\infty}^{+\infty} \frac{F[\dot{\varphi}_x(\tau)]}{2\pi} d\omega$$

Dimostrazione pag. 125

La funzione integranda al secondo membro prende il nome di densità spettrale di energia. Si noti che essa lega, mediante trasformata di Fourier, la funzione di autocorrelazione allo spettro di energia.

$$E_{bil}(\omega) = \frac{F[\dot{\varphi}_x(\tau)]}{2\pi} = \frac{|X(\omega)|^2}{2\pi}$$

Se $x(t)$ è reale si può definire una densità spettrale di energia monolatera come

$$E(\omega) = \begin{cases} 2E_{bil}(\omega) & \omega \geq 0 \\ E_{bil}(\omega) & \omega = 0 \end{cases}$$

Le componenti spettrali di $x(t)$ all'interno di un certo intervallo hanno energia data dall'integrale della densità spettrale sul medesimo intervallo:

$$E_{\omega_1, \omega_2} = \int_{\omega_1}^{\omega_2} E(\omega) d\omega$$

Densità di energia riferita alla frequenza

Se si preferisce riferirsi alle frequenze anziché alle pulsazioni si ha

$$E_{f,bil}(f) = 2\pi E_{bil}(2\pi f)$$

per la quale

$$E = \int_{-\infty}^{+\infty} E_{f,bil}(f) df$$

Segnali a potenza finita

Funzioni di crosscorrelazione ed autocorrelazione

Le definizioni delle funzioni di crosscorrelazione ed autocorrelazione per le funzioni a potenza finita si possono ottenere con la seguente sostituzione formale:

$$\int_{-\infty}^{+\infty} dt \rightarrow \lim_{T \rightarrow \infty} \frac{1}{T} \int_{-\frac{T}{2}}^{+\frac{T}{2}} dt$$

Si ha quindi per la funzione di crosscorrelazione:

$$\varphi_{xy}(\tau) = \lim_{T \rightarrow \infty} \frac{1}{T} \int_{-\frac{T}{2}}^{+\frac{T}{2}} x^*(t)y(t + \tau)dt$$

e quella di autocorrelazione

$$\varphi_x(\tau) = \lim_{T \rightarrow \infty} \frac{1}{T} \int_{-\frac{T}{2}}^{+\frac{T}{2}} x^*(t)x(t + \tau)dt$$

Proprietà delle funzioni di cross e autocorrelazione

Per le funzioni appena definite valgono le seguenti proprietà, analoghe a quelle viste per i segnali ad energia finita:

$$\begin{aligned}\varphi_{xy}(\tau) &= \varphi_{yx}^*(-\tau) \\ \varphi_x(\tau) &= \varphi_x^*(-\tau)\end{aligned}$$

$$\begin{aligned}|\varphi_{xy}|^2 &\leq \varphi_x(0)\varphi_y(0) = P_x P_y \\ |\varphi_x(\tau)| &\leq \varphi_x(0) = P_x\end{aligned}$$

Densità spettrale di potenza

Analogamente a quanto visto in precedenza per le funzioni ad energia finita si ottiene:

$$P = \int_{-\infty}^{+\infty} \frac{F[\varphi_x(\tau)]}{2\pi} d\omega$$

Dimostrazione pag. 126

La funzione integranda prende il nome di densità spettrale di potenza:

$$G_{bil}(\omega) = \frac{F[\varphi_x(\tau)]}{2\pi}$$

In modo del tutto analogo a quanto visto in precedenza, si può ottenere la densità spettrale riferita alle frequenze

$$F_{f,bil}(f) = 2\pi G_{bil}(2\pi f)$$

Anche nel caso di funzioni a potenza finita, la denominazione di spettro di potenza è giustificata dalla proprietà locale per cui:

$$P_{\omega_1, \omega_2} = \int_{\omega_1}^{\omega_2} G(\omega) d\omega$$

Segnali a potenza finita periodici

I segnali periodici rappresentano un caso particolare di segnali a potenza finita, in quanto per essi è possibile avere altre rappresentazioni nel dominio delle frequenze (sviluppi in serie di Fourier, trasformata generalizzata). Risulta interessante individuare comunque anche il loro spettro di potenza, per collegarlo alle altre rappresentazioni.

Si noti innanzitutto che le funzioni di cross ed autocorrelazione di segnali periodici di ugual periodo T , possono essere calcolate su un periodo.

Applicando quindi la seguente sostituzione formale

$$\lim_{T \rightarrow \infty} \frac{1}{T} \int_{-\frac{T}{2}}^{+\frac{T}{2}} dt \rightarrow \frac{1}{T} \int_{-\frac{T}{2}}^{+\frac{T}{2}} dt$$

si ottengono le definizioni di funzione di crosscorrelazione

$$\varphi_{xy}(\tau) = \frac{1}{T} \int_{-\frac{T}{2}}^{+\frac{T}{2}} x^*(t)y(t + \tau) dt$$

e quella di autocorrelazione

$$\varphi_x(\tau) = \frac{1}{T} \int_{-\frac{T}{2}}^{+\frac{T}{2}} x^*(t)x(t + \tau) dt$$

Per individuare lo spettro di potenza è possibile dimostrare che la funzione di autocorrelazione è anch'essa una funzione periodica dello stesso periodo T , ovvero

$$\varphi_x(\tau) = \sum_{n=-\infty}^{\infty} c_n^* c_n e^{in\omega_0 \tau} = \sum_{n=-\infty}^{\infty} |c_n|^2 e^{jn\omega_0 \tau}$$

Dimostrazione pag. 127

Dall'espressione sopra si ricava immediatamente la densità spettrale di potenza,

$$G_{bil}(\omega) = \frac{F[\varphi_x(\tau)]}{2\pi} = \sum_{n=-\infty}^{\infty} |c_n|^2 \delta(\omega - n\omega_0)$$

Nel caso di segnali reali essa può anche essere scritta in forma monolatera utilizzando le relazioni che legano i coefficienti A_n e c_n :

$$G(\omega) = A_0^2 \delta(\omega) + \sum_{n=1}^{+\infty} \frac{A_n^2}{2} \delta(\omega - n\omega_0)$$

Trasformazioni lineari tempo invarianti di spettri di energia e di potenza

Si supponga che i segnali $y(t)$ ed $x(t)$ si trovino rispettivamente in ingresso ed in uscita ad una rete lineare tempo invariante, avente risposta impulsiva $h(t)$ e funzione di trasferimento $H(\omega)$. Gli spettri di energia o di potenza dei due segnali sono legati dal modulo della $H(\omega)$ al quadrato, ma per dimostrarlo conviene separare i casi.

Segnali ad energia finita

$$E_{y,bil}(\omega) = |H(\omega)|^2 E_{x,bil}(\omega)$$

Dimostrazione

Ricordando che

$$Y(\omega) = H(\omega)X(\omega)$$

si ha

$$\frac{|Y(\omega)|^2}{2\pi} = |H(\omega)|^2 \frac{|X(\omega)|^2}{2\pi}$$

QED.

Segnali a potenza finita

$$G_{y,bil}(\omega) = |H(\omega)|^2 G_{x,bil}(\omega)$$

Dimostrazione pag. 129

Esercizi

Energia e potenza di un segnale tempo discreto

Per i segnali tempo discreti si può procedere come nel caso dei segnali tempo continui, con i necessari cambi formali. Sotto ci si limiterà ad alcune considerazioni essenziali, utili in seguito.

Definita la potenza istantanea come $|x_n|^2$, derivano le seguenti definizioni di "energia" e "potenza media", o semplicemente potenza:

$$E = \sum_{n=-\infty}^{+\infty} |x_n|^2$$
$$P = \langle |x_n|^2 \rangle = \lim_{N \rightarrow \infty} \frac{1}{2N+1} \sum_{n=-N}^{+N} |x_n|^2$$

Per le serie ad energia finita esiste la trasformata di Fourier. La funzione di autocorrelazione ha l'espressione:

$$\dot{c}_k = \sum_{n=-\infty}^{+\infty} x_n^* x_{n+k}$$

Per le serie a potenza finita non esiste di norma la trasformata di Fourier. La funzione di autocorrelazione è data dalla seguente media temporale:

$$c_k = \langle x_n^* x_{n+k} \rangle = \lim_{N \rightarrow \infty} \frac{1}{2N+1} \sum_{n=-N}^{+N} x_n^* x_{n+k}$$

▼ 9.0 - Cenni sui segnali digitali

▼ 9.1 - Spettri di segnali PAM

Spettri di segnali PAM deterministici

I segnali PAM (Pulse Amplitude Modulation) sono esprimibili come convoluzione fra una serie temporale $\{a_n\}$ ed un impulso ad energia finita $g(t)$:

$$s(t) = \sum_{n=-\infty}^{\infty} a_n g(t - nT) = \{a_n\}^* g(t)$$

Il segnale PAM può essere analogico, se la serie è analogica, digitale se invece può assumere solo un numero finito di valori. Se la serie temporale è ad energia finita, allora anche il segnale PAM è ad energia finita, e la sua trasformata è data dal prodotto delle trasformate:

$$S(\omega) = A_s(\omega)G(\omega)$$

Se invece la serie temporale è a potenza finita, il segnale PAM è anch'esso a potenza finita e non è trasformabile secondo Fourier, per cui occorre passare ad un'analisi generalizzata, ovvero calcolare il suo spettro di potenza. Se ne daranno qui i soli passi essenziali.

Per prima cosa assumiamo la serie a valore medio nullo; si può dimostrare che la sua funzione di autocorrelazione equivale a:

$$\varphi_s(\tau) = \frac{1}{T} \{c_k\}^* \dot{\varphi}_g(\tau)$$

Come per tutti i segnali a potenza finita, lo spettro di potenza del segnale PAM è dato dalla trasformata della funzione di autocorrelazione:

$$G_{s,bil}(\omega) = \frac{\varphi_s(\tau)}{2\pi} = \frac{1}{2\pi T} F[\dot{\varphi}_g(\tau)] F[\{c_k\}] = \frac{|G(\omega)|^2}{2\pi T} \sum_{k=-\infty}^{\infty} c_k e^{-jk\omega T}$$

Se la serie temporale $\{a_n\}$ è reale, la sua autocorrelazione $\{c_n\}$ è reale e pari, per cui la sua trasformata può essere scritta con solo riferimento agli indici positivi:

$$\begin{aligned} \sum_{k=-\infty}^{\infty} c_k e^{-jk\omega T} &= c_0 + 2 \sum_{k=1}^{\infty} c_k \cos k\omega T \\ \implies G_{s,bil} &= \frac{|G(\omega)|^2}{2\pi T} \sum_{k=-\infty}^{\infty} [c_0 + 2 \sum_{k=1}^{\infty} c_k \cos k\omega T] \end{aligned}$$

Infine se la autocorrelazione è nulla per k diverso da zero, si ottiene:

$$G_{s,bil}(\omega) = \frac{c_0 |G(\omega)|^2}{2\pi T}$$

Spettri di segnali PAM aleatori

Un segnale è aleatorio quando il suo andamento nel tempo è aleatorio, ovvero è una funzione aleatoria. Un segnale PAM è aleatorio se lo è la serie temporale $\{a_n\}$. Per

una serie aleatoria non è nota a priori la successione dei valori; tuttavia ciò non significa che la conoscenza della serie sia nulla, infatti la serie può sempre essere descritta in termini statistici, come successione di infinite variabili aleatorie; in particolare se la serie è stazionaria tutte le valutazioni statistiche sono indipendenti dall'indice n che denota la posizione della

variabile all'interno della sequenza. A titolo di esempio, si riporta la formula del valore medio statistico di una serie stazionaria:

$$E[a_n] = \sum_{i=1}^L a^i P(a^i)$$

dove gli a^i rappresentano i valori che possono essere assunti e le $P(a^i)$ le rispettive probabilità.

Per le serie aleatorie è possibile definire una autocorrelazione statistica (a priori) come media statistica del prodotto delle coppie di valori posti a distanza k , ovvero come $E[a_n^* a_{n+k}]$; se la serie è stazionaria, come supporremo d'ora in avanti, la probabilità della coppia $P(a_i, a_l, k)$ non dipende dalla posizione delle due variabili aleatorie all'interno della sequenza, cioè dal pedice "n" che rappresenta la posizione del primo elemento, ma solo dalla distanza fra gli elementi, cioè da "k":

$$c_{stat,k} = E[a_n^* a_{n+k}] = \sum_{i=1}^L \sum_{l=1}^L (a^i)^* a^l P(a_i, a_l, k)$$

Le variabili aleatorie a distanza k sono incorrelate se:

$$c_{stat,k} = E[a_n^* a_{n+k}] = \begin{cases} E[|a_n|^2] & k = 0 \\ E[a_n^*] E[a_{n+k}] = E[a_n^*] E[a_n] & k \neq 0 \end{cases}$$

Condizione sufficiente per l'incorrelazione è che le variabili aleatorie a_n e a_{n+k} siano indipendenti.

Se inoltre il valor medio è nullo le variabili sono incorrelate se e solo se l'autocorrelazione è nulla per k diverso da zero.

Infine se la serie è "ergodica" (è sufficiente che oltre che stazionaria sia a memoria finita, cioè la variabili a_n e a_{n+k} siano indipendenti per valori di k sufficientemente grandi) i valori medi statistici coincidono con gli analoghi temporali che possono essere calcolati a posteriori sulle realizzazioni del processo stocastico, cioè sulle serie temporali che si ottengono a posteriori, come risultato dell'esperimento. Si ha quindi in particolare che la autocorrelazione statistica e quella temporale vengono a coincidere per i processi ergodici:

$$c_{stat,k} = c_k$$

Si hanno una serie di conseguenze a catena:

- tutte le realizzazioni di un processo ergodico, anche se ovviamente diverse fra loro, hanno la stessa autocorrelazione temporale, dovendo questa coincidere con quella statistica, che è unica.
- quanto sopra implica che tutte le realizzazioni dei segnali PAM, pur diverse fra loro in quanto è diversa la successione dei simboli della serie, hanno lo stesso spettro di potenza; esso si ottiene dalla formula per i segnali deterministicamente sostituendo alla autocorrelazione temporale quella statistica.

In particolare, se la serie è a valor medio nullo e gli elementi della serie aleatoria sono incorrelati si ottiene infine:

$$G_{s,bil}(\omega) = \frac{E[|a_n|^2]|G(\omega)|^2}{2\pi T}$$

▼ 9.2 - Cenni sui segnali digitali aleatori in banda base

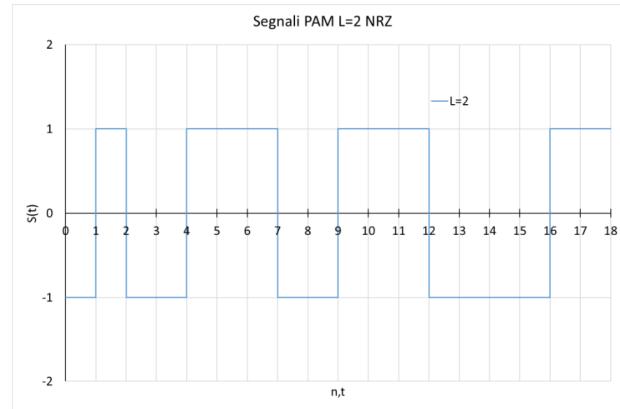
Un sistema di trasmissione digitale in banda base è normalmente modellato come una sorgente binaria, i cui simboli, o bit, sono equiprobabili e indipendenti, seguita da un codificatore di linea che trasforma la serie di bit $\{b_n\}$ in una serie di simboli digitali, $\{a_n\}$. Il segnale numerico che viene trasmesso è il segnale PAM che si ottiene convolvendo questa serie temporale con l'impulso $g(t)$. Vedremo qui due esempi molto semplici di codifica.

Codifica binaria

Questa codifica si limita a trasformare i bit a "0" in "-1", allo scopo di rendere a valor medio nullo la serie temporale degli $\{a_n\}$.

Dato che ho un simbolo per ogni bit, allora il tempo di simbolo, cioè l'intervallo fra simboli T coincide con il tempo di bit T_b , ed analogamente per i loro inversi, cioè le frequenze di simbolo f_s e di bit, f_b . Dato che i bit sono indipendenti, lo sono per forza anche i simboli a_n , da cui deriva che:

$$c_{stat,k} = \begin{cases} E[a_n^2] = 1 & k = 0 \\ E[a_n]E[a_n] = 0 & k \neq 0 \end{cases}$$



Codifica multilivello

La codifica multilivello è un'estensione della codifica bipolare, nella quale viene emesso un simbolo ogni l bit, per cui si ha:

$$T = lT_b$$

$$f_s = \frac{f_b}{l}$$

I simboli possono assumere i seguenti $L = 2^l$ valori: $a^i = \pm 1, \pm 3, \dots, \pm (L - 1)$.

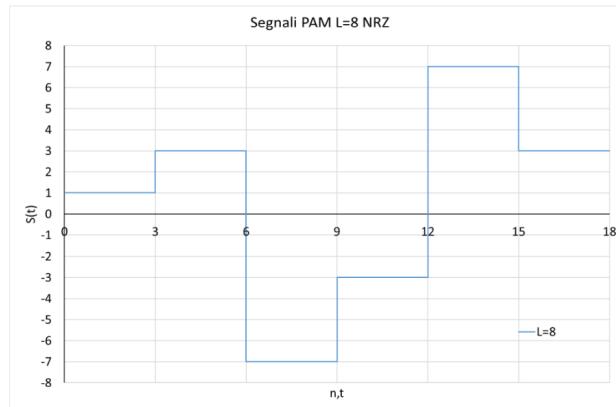
La scelta dei numeri dispari è data dal fatto che si vuole che ogni valore abbia la stessa distanza, cioè due, dal valore successivo. Anche se non è obbligatorio, conviene mappare gruppi di l bit in un simbolo adottando una codifica di Gray:

<i>l-pla</i>	Simbolo
000	+7
001	+5
011	+3
010	+1
110	-1
111	-3
101	-5
100	-7

I simboli a_n sono indipendenti ed equiprobabili, in quanto ottenuti da gruppi di l bit distinti (senza bit in comune). Il valor medio è quindi nullo mentre la funzione di autocorrelazione statistica è data da:

$$c_{stat,k} = \begin{cases} E[a_n^2] = \frac{L^2 - 1}{3} & k = 0 \\ E[a_n]E[a_n] = 0 & k \neq 0 \end{cases}$$

L'espressione della potenza statistica $E[a_n^2] = \frac{L^2 - 1}{3}$ non è stata dimostrata.



Spettri dei segnali PAM aleatori con codifica multilivello e impulso rettangolare NRZ

Nel caso di codifica multilivello (comprendente anche il caso bipolare), la autocorrelazione è nulla tranne che nell'origine. Lo spettro del segnale PAM diventa quindi:

$$G_{s,bil}(\omega) = \frac{E[a_n^2]|G(\omega)|^2}{2\pi T}$$

dove $E[a_n^2] = \frac{L^2 - 1}{3}$.

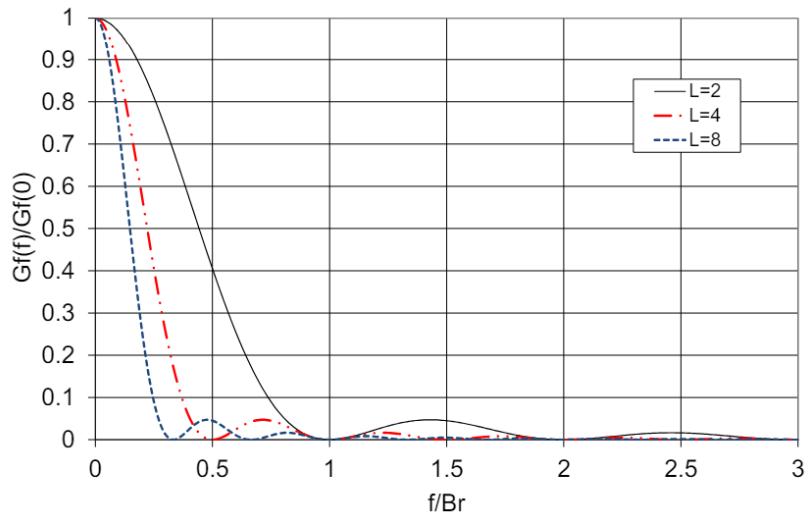
Come impulso g(t) si può prendere l'impulso rettangolare di durata T (detto anche NRZ, No Return to Zero). La sua trasformata è data da: $G(\omega) = T \frac{\sin \omega T / 2}{\omega T / 2}$. Da cui:

$$G_{s,bil}(\omega) = \frac{E[a_n^2]T}{2\pi} \left| \frac{\sin \omega T / 2}{\omega T / 2} \right|^2$$

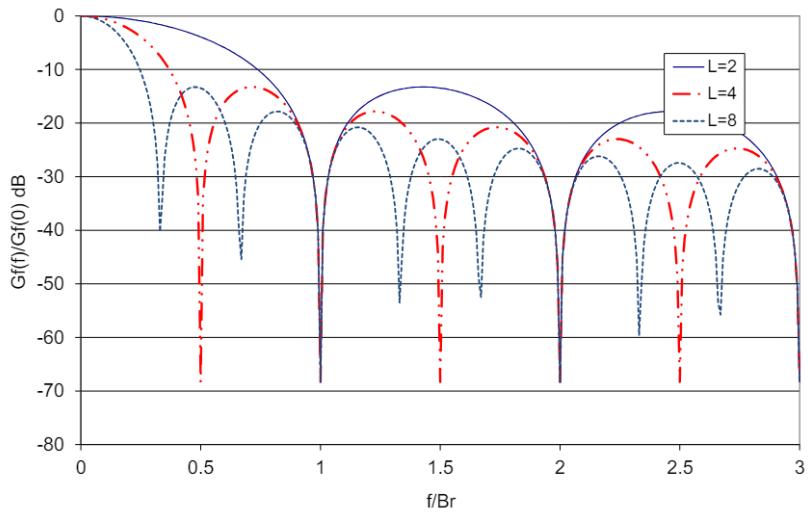
Conviene passare alle frequenze:

$$G_{s,f,bil}(f) = E[a_n^2]T \left| \frac{\sin 2\pi f T / 2}{2\pi f T / 2} \right|^2 = E[a_n^2]T |\text{sinc } f T|^2 = \frac{E[a_n^2]}{f_s} \left| \text{sinc } \frac{f}{f_s} \right|^2$$

Si noti che la frequenza di simbolo coincide con la frequenza di bit nel caso bipolare, cioè quando $I=1$, altrimenti è ridotta di un fattore I . Il vantaggio della codifica multilivello risiede nel fatto che la banda del primo lobo (coincidente con il primo punto di nullo) è pari alla frequenza di simbolo, per cui, a parità di frequenza di bit, essa risulta ridotta del medesimo fattore I . Ciò risulta evidente osservando la figura sotto:



Spettri di potenza al variare del numero di livelli.



Analoga alla figura precedente, ma con la scala in dB (i punti di nullo vanno a $-\infty$).

Si può dimostrare che il prezzo da pagare per la riduzione di banda è dato da un aumento della probabilità di errore per bit, che aumenta all'aumentare di L . Le codifiche multilivello, con L sempre più elevato quindi si impiegano quando è necessario risparmiare banda e le caratteristiche di ricezione (rapporto segnale rumore) sono buone.

Cenni sulle modulazioni digitali (segnali digitali aleatori passa banda)

In linea di principio le modulazioni digitali si ottengono dalle modulazioni analogiche sostituendo al segnale modulante analogico un segnale modulante digitale, cioè di solito un segnale PAM a L livelli con codifica NRZ.

Esistono però modulazioni analogiche che non hanno un corrispondente digitale (l'AM), e viceversa, come l'FSK. Nella tabella sotto vengono mostrate alcune corrispondenze:

Analogica	Digitale
AM, o DSB	
Prodotto o DSB-SC	ASK, L-ASK
QAM	M-QAM, QPSK
PM	PSK, BPSK, 2-PSK, L-PSK
FM	L-CP-FSK, (MSK)
	FSK, L-FSK

Il segnale modulato $s(t)$ è a potenza finita e per le modulazioni ASK, PSK, M-QAM il suo spettro di potenza $G_{s,bil}(\omega)$ si ottiene dallo spettro di potenza $G_{x,bil}(\omega)$ del segnale modulante in modo analogo a quanto visto per la trasformata di Fourier di una modulazione a prodotto, con l'unica differenza della costante 1/4 anziché 1/2.

$$G_{s,bil}(\omega) = \frac{1}{4} G_{x,bil}(\omega - \omega_o) + \frac{1}{4} G_{x,bil}(\omega + \omega_o)$$

Si ha anche in questo caso, come nella modulazione a prodotto, un raddoppio della banda rispetto a quella del segnale modulante. In realtà, per occupare meno banda radiofrequenza, il segnale modulante viene opportunamente filtrato passabasso, in modo che la sua frequenza massima sia compresa fra la metà della frequenza di simbolo $f_s/2$ e la frequenza di simbolo stessa f_s . Di conseguenza l'oscillazione modulata corrispondente ha una banda compresa fra la frequenza di simbolo f_s e $2f_s$.

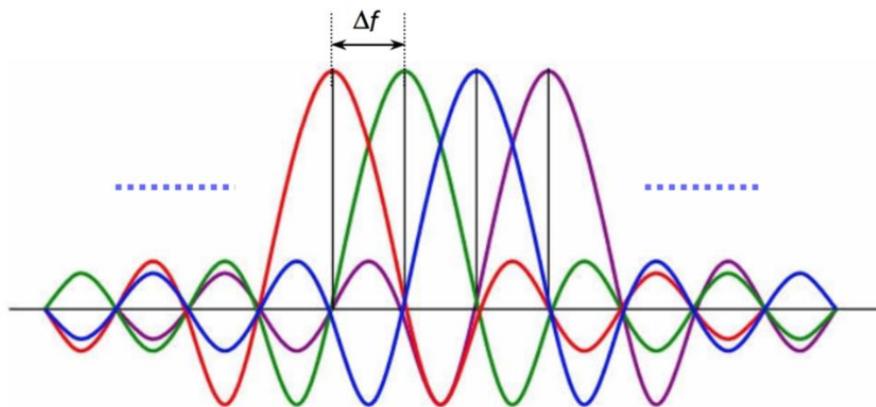
Modulazioni digitali utilizzate nell'802.11n (WiFi)

Lo standard 802.11 n, il "routerino" WiFi di casa, utilizza una tecnica multiportante detta OFDM (Orthogonal Frequency-Division Multiplexing). Senza entrare nel merito dell'OFDM, quello che interessa rilevare qui è che per ogni valore dei flussi spaziali (da 1 a 4) si hanno 8 combinazioni modulazione-codifica. Le modulazioni sono protette da codici a correzione di errore che introducono ridondanza. Ad esempio, un codice con "coding rate" $\frac{3}{4}$ inserisce un bit di ridondanza ogni 3 di informazione.

La frequenza di bit che si ottiene dipende dall'indice MCS, e per ogni MCS dall'ampiezza di banda a disposizione e dal valore scelto per il tempo di guardia. Si va da 6.5 a 600 Mbit/s, cioè due ordini di grandezza! Il router seleziona automaticamente la combinazione migliore in base al numero delle antenne ed al rapporto segnale/rumore del canale. Naturalmente la frequenza di bit sarà più alta se il canale è migliore. In ogni caso è altamente probabile nei casi domestici che il fattore limitante per l'accesso a Internet sia la banda della connessione ADSL. Lo standard "n" può operare nella banda non licenziata a 2.4 GHz, oppure, facoltativamente, su quella a 5 GHz.

Tempo utile e tempo di simbolo

L'OFDM divide una banda B in N sottoportanti equidistanziate di $\Delta f = B/N$. N è la dimensione della FFT. Le portanti sono ortogonali su $T = 1/\Delta f$ perché su T ognuna di esse ha un numero intero di periodi. Si usa quindi T come tempo utile di simbolo, o T_u . Al tempo utile viene aggiunto un tempo di guardia T_g per permettere ai cammini multipli di esaurirsi evitando così l'interferenza intersimbolo, ottenendo così il tempo di simbolo OFDM, T_{OFDM} . Ogni portante sarà modulata a L livelli quindi porterà m bit per ogni tempo di simbolo OFDM. L'inverso del tempo di simbolo OFDM fornisce la frequenza dei simboli OFDM.



Spaziatura delle sottoportanti di un segnale OFDM.

▼ 9.3 - Misure in decibel

Le misure in decibel si riferiscono a rapporti di grandezze omogenee, ed in quanto tali sono adimensionali. Sono misure logaritmiche, particolarmente adatte a rappresentare grandezze molto diverse fra loro, come spesso accade nei circuiti elettrici. Ad esempio il guadagno di potenza di un doppio bipolo può essere espresso in decibel (dB) in questo modo:

$$G_{p,dB} = 10 \log_{10} \left(\frac{P_y}{P_x} \right)$$

Dalle proprietà dei logaritmi discende che il prodotto di due guadagni diventa la somma dei medesimi una volta espressi in decibel, e che la divisione per un guadagno diventa una sottrazione:

$$\begin{aligned} G &= G_1 G_2 & G_{dB} &= G_{1,dB} + G_{2,dB} \\ G &= \frac{G_1}{G_2} & G_{dB} &= G_{1,dB} - G_{2,dB} \end{aligned}$$

Ad ogni ordine di grandezza in più corrisponde un incremento di 10 dB, e viceversa.

La definizione e le formule riportate sopra si possono applicare in generale a qualsiasi rapporto di grandezze. Tuttavia, in ambito ingegneristico, per la misura di rapporti di grandezze legate alla radice quadrata di potenze, cioè in pratica a rapporti di tensioni, correnti, intensità di campo, viene usata una definizione leggermente diversa:

$$G_{v,dB} = 20 \log_{10} \left(\frac{V_y}{V_x} \right)$$

La ragione è legata al desiderio ingegneristico di non dover specificare se il guadagno è in tensione (o corrente) o in potenza. Infatti, ripartendo dal guadagno di potenza, supponendo resistive le impedenze di ingresso e di carico di un doppio bipolo si ha:

$$G_{p,dB} = 10 \log_{10} \left(\frac{P_y}{P_x} \right) = 10 \log_{10} \left(\frac{V_y^2 / R_c}{V_x^2 / R_i} \right)$$

Nel caso frequente in cui la resistenza di ingresso e quella di uscita coincidano, dalla formula sopra deriva la coincidenza del guadagno di potenza in dB con quello di tensione in dB:

$$G_{p,dB} = 20 \log_{10} \left(\frac{V_y}{V_x} \right) = G_{v,dB}$$

Come detto le misure in decibel sono adimensionali, riferendosi a rapporti di grandezze omogenee. A volte tuttavia è comodo utilizzare i decibel per rappresentare potenze e tensioni, anziché rapporti delle medesime. In

questo caso si procede considerando il rapporto fra la grandezza, potenza o tensione, e la sua unità di misura, che ora viene indicata nel pedice:

$$P_{dBW} = 10 \log_{10}\left(\frac{P}{1W}\right)$$
$$P_{dBV} = 20 \log_{10}\left(\frac{X}{1V}\right)$$

▼ 10.0 - Internet

▼ 10.1 - Introduzione a Internet

Cenni storici su Internet

- 1957

Viene lanciato il primo satellite artificiale, lo Sputnik. Non è geostazionario e non è un satellite per telecomunicazioni. Ha però una radio a bordo che emette dei bip, ascoltabili dai radioamatori di tutto il mondo. È figlio della Guerra Fredda. Inizia la corsa allo spazio.

Il razzo USA Vanguard Kaputnik con a bordo il satellite VT3 esplode sulla rampa di lancio in diretta TV. Alla perdita del primato si aggiunge l'umiliazione del fallimento. La stampa USA è feroce.

- 1958

Nel febbraio 1958 viene fondata l'ARPA (Advanced Research Projects Agency, poi DARPA) Lo scopo è quello di assicurare la supremazia tecnologica degli Stati Uniti. Luglio 1958: Fondata la NASA (National Aeronautics and Space Administration).

- Anni 60

Paul Baran (RAND Co.) iniziò ad interessarsi alla possibilità di realizzare una rete di telecomunicazioni in grado di sopravvivere ad un attacco nucleare. Nascita di Internet. Nel sistema telefonico tradizionale, a causa della struttura gerarchica, fra A e B un solo percorso è possibile.

Basi di Internet

Elementi essenziali del progetto:

- Architettura distribuita (non un unico punto di fallimento) e ridondante (più percorsi)
- Comutazione di pacchetto (Packet switching) di tipo «connectionless» al posto di commutazione di circuito (Circuit switching)

La commutazione di pacchetto divide i messaggi in pacchetti di lunghezza arbitraria; se "connectionless" ogni pacchetto è instradato autonomamente come una lettera.

Dagli anni 60 al 1991 succede: creazione ARPANET, protocolli TCP/IP e creazione World Wide Web.

Modelli di rete

Ci sono due tipi di tecnologie di trasmissione diffusamente impiegati: i collegamenti broadcast e i collegamenti punto a punto (point to point). I collegamenti punto a punto connettono coppie di computer. I pacchetti possono dover visitare una o più macchine intermedie per spostarsi dalla sorgente alla destinazione in una rete composta di collegamenti punto a punto. Spesso sono possibili più percorsi di diversa lunghezza; quindi, nelle reti punto a punto è importante trovarne di validi.

Al contrario, le reti broadcast hanno un solo canale di comunicazione condiviso da tutte le macchine della rete. I pacchetti inviati da qualunque macchina sono ricevuti da tutte le altre. Un campo indirizzo all'interno del pacchetto individua il destinatario. Alla ricezione del pacchetto ogni macchina controlla il campo indirizzo. Se il pacchetto è indirizzato alla macchina ricevente viene processato; se è indirizzato a un'altra macchina viene semplicemente ignorato. I sistemi broadcast danno di solito anche la possibilità d'indirizzare un pacchetto a tutti i destinatari.

Un criterio alternativo per classificare le reti è la loro scala. La distanza è una metrica di classificazione importante, perché reti su scale differenti impiegano tecnologie diverse.

- PAN (Personal Area Network): rete molto piccola ristretta a pochi pc
- LAN (Local Area Network): rete locale la cui dimensione può essere relativa a un edificio o un campus, quindi di dimensioni geografiche limitate.

Le LAN possono usare una tecnologia di trasmissione rappresentata da un cavo a cui sono connesse tutte le macchine. Per le LAN broadcast sono possibili differenti topologie (es. bus e anello).

Le reti broadcast si possono ulteriormente dividere in statiche e dinamiche, a seconda del modo in cui è allocato il canale. Una tipica allocazione statica consiste nel suddividere il tempo in intervalli discreti, permettendo a ogni macchina di eseguire il broadcast solo quando è attivo il proprio turno. I metodi di allocazione dinamica per un canale condiviso possono essere centralizzati o non centralizzati. Nel metodo di

allocazione centralizzato esiste una singola entità, che stabilisce a chi spetta di volta in volta l'uso del mezzo. Nel metodo di allocazione non centralizzato non esiste un'entità centrale, ogni macchina deve decidere in autonomia se trasmettere.

- MAN (Metropolitan Area Network): Rete che copre un intera città
- WAN (Wide Area Network): rete che copre un'intero stato o continente
Racchiude una raccolta di macchine destinate a eseguire programmi utente, chiamate host. Gli host sono collegati tra loro tramite una sottorete.

Le WAN, come le abbiamo descritte, sembrano simili a una LAN cablata su larga scala, ma vi sono importanti differenze che vanno al di là della lunghezza dei cavi. Ad esempio i router connettono usualmente reti diverse dal punto di vista tecnologico (es. ethernet, sonet) e di tipo (singoli computer o intere reti LAN).

- Internet: copre l'intero pianeta

Esistono anche diverse tipologie di reti wireless:

- Reti interconnesse:

Un insieme di reti interconnesse si chiama internetwork o internet (con la *i* minuscola). Questi termini sono usati nel loro significato generico, in contrasto con la Internet mondiale (che è una specifica internet) che indicheremo sempre con l'iniziale maiuscola. La internet mondiale usa le reti ISP per connettere reti aziendali, reti domestiche e molte altre reti.

- LAN wireless:

Le reti LAN wireless sono attualmente molto diffuse. In questi sistemi ogni computer ha un ricevitore radio e un'antenna per comunicare con altri computer. Nella maggior parte dei casi ogni computer comunica con un dispositivo posto sul soffitto, chiamato AP (access point), router wireless, o base station. Tuttavia, se gli altri computer sono abbastanza vicini tra loro, possono comunicare direttamente in una configurazione peer-to-peer. Lo standard per le reti LAN wireless è IEEE 802.11, comunemente noto come Wi-Fi e attualmente molto diffuso.

- WAN wireless:

Alcuni tipi di reti WAN fanno uso di tecnologie wireless: nei sistemi satellitari ogni computer a terra ha un'antenna attraverso la quale può trasmettere e ricevere dati da un satellite in orbita. Tutti i computer possono ascoltare messaggi provenienti dal satellite e in alcuni casi

anche le trasmissioni trasmesse da terra da altri computer verso il satellite. Le reti satellitari sono intrinsecamente broadcast.

Un altro esempio di WAN wireless è la rete telefonica cellulare.

Software di rete

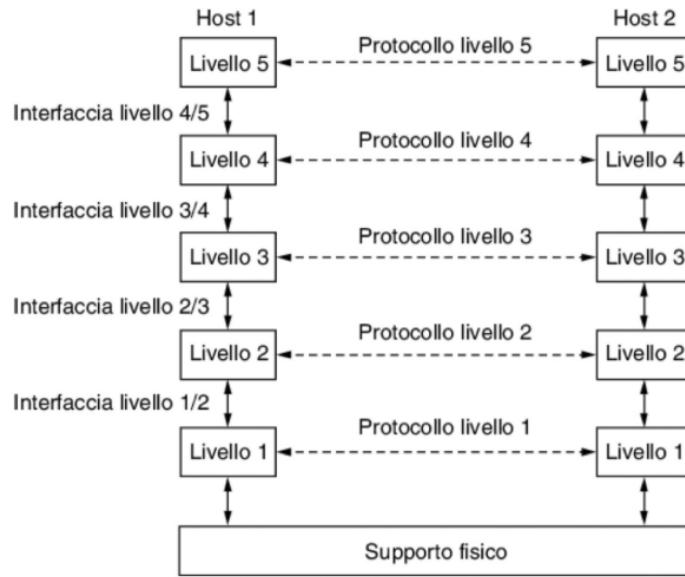
Gerarchie dei protocolli

Per diminuire la complessità, la maggior parte delle reti è organizzata come una pila di livelli (layer) o strati, costruiti uno sull'altro. Lo scopo di ogni livello è quello di offrire determinati servizi ai livelli superiori, schermandola dai dettagli implementativi.

Quando il livello n all'interno di un computer è in comunicazione con il livello n di un altro computer, le regole e le convenzioni usate in questa comunicazione sono globalmente note come protocolli del livello n.

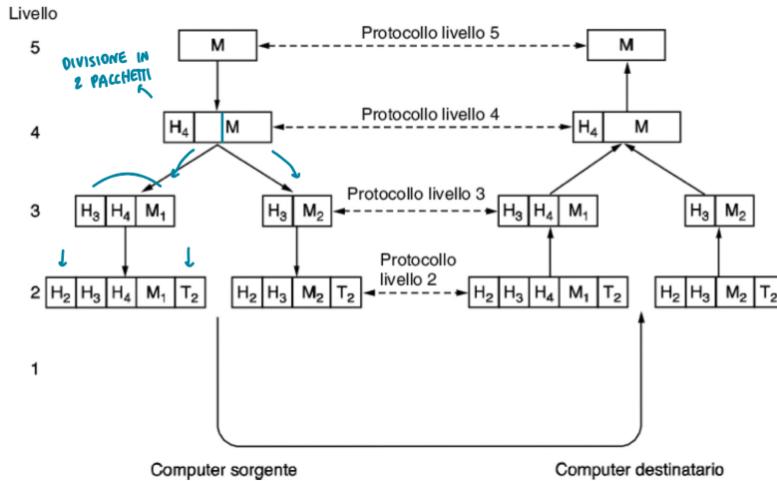
Fondamentalmente, un protocollo è un accordo tra le parti che comunicano sul modo in cui deve procedere la comunicazione. Le entità che formano i livelli di pari grado su diversi computer sono chiamati peer (pari). I peer possono essere processi software, dispositivi hardware o esseri umani. Sono i peer a comunicare tra loro usando il protocollo.

In realtà i dati non sono trasferiti direttamente dal livello n di un computer al livello n di un altro. Ogni livello passa i dati a quello sottostante, fino a raggiungere il più basso. Sotto al livello 1 si trova il supporto fisico attraverso cui è possibile la comunicazione vera e propria. Tra ciascuna coppia di livelli contigui si trova un'interfaccia, che definisce le operazioni elementari e i servizi che il livello inferiore rende disponibili a quello soprastante. L'insieme di livelli e protocolli si chiama architettura di rete. L'elenco dei protocolli usati da uno specifico sistema è chiamato pila di protocolli (protocol stack).



Consideriamo un esempio: un messaggio M è prodotto da un processo applicativo che lavora al livello 5 e passato al livello 4 per la trasmissione. Il livello 4 mette un header (intestazione) davanti al messaggio per identificarlo e passa il risultato al livello 3. L'header include informazioni di controllo, come l'indirizzo, per consentire al livello 4 del computer destinatario di consegnare il messaggio.

In molte reti non c'è un limite alla dimensione dei messaggi trasmessi nel protocollo del livello 4, ma quasi sempre c'è un limite imposto dal protocollo del livello 3. Di conseguenza, il livello 3 deve spezzare i messaggi in arrivo in unità più piccole chiamate pacchetti, aggiungendo un header di livello 3 davanti ad ogni pacchetto. In questo esempio M è diviso in due parti, M1 ed M2 che vengono trasmessi separatamente. Il livello 3 decide la linea di uscita da usare e passa i pacchetti al livello 2. Il livello 2 aggiunge a ogni pezzo un trailer (informazione aggiuntiva in coda) e lo passa al livello 1 per la trasmissione fisica. Nel computer destinatario il messaggio si muove verso l'alto, passando da livello in livello, e gli header vengono rimossi man mano.



Progettazione dei livelli

Alcuni problemi fondamentali nella progettazione delle reti si presentano livello dopo livello:

1. Controllo degli errori: individuazione di errori (error detection) e correzione degli errori (error correction).
2. Routing (instradamento): consiste nel trovare un percorso valido attraverso la rete, che dovrebbe prendere questa decisione in modo automatico.
3. Addressing o naming: ogni livello richiede un meccanismo per identificare chi trasmette e chi riceve un particolare messaggio.
4. Multiplexing: molti progetti di rete condividono dinamicamente la banda concedendola alle necessità a breve termine degli host, piuttosto che assegnare a ogni host una frazione prefissa di banda che ciascuno può o meno usare.
5. Flow control (controllo di flusso): è un feedback da parte del ricevente alla sorgente, utilizzato nell'eventualità per impedire che una sorgente veloce inondi di dati un ricevente lento. A volte la rete è intasata perché troppi computer cercano di inviare un traffico eccessivo di dati e non riesce a consegnare i pacchetti. Tale sovraccarico della rete è chiamato congestione.

Servizi orientati alla connessione e senza connessione

Ogni livello può offrire a quelli sovrastanti due tipi diversi di servizio: orientati alla connessione oppure senza connessione.

Un servizio orientato alla connessione assomiglia al sistema telefonico.

L'utente deve innanzitutto stabilire una connessione, usarla e quindi

rilasciarla. L'aspetto essenziale di una connessione è che funziona come un tubo: il trasmettitore vi spinge oggetti (bit) a una estremità e il ricevitore li prende dall'altra.

Al contrario del servizio orientato alla connessione, un servizio senza connessione si comporta come il servizio postale. Ogni messaggio (pacchetto nel livello rete) porta l'indirizzo completo del destinatario ed è instradato attraverso il sistema postale in modo indipendente dai messaggi successivi.

Ogni tipo di servizio può ulteriormente essere caratterizzato in base alla sua affidabilità. Alcuni servizi sono affidabili, nel senso che non perdono mai dati.

Primitive di servizio

Un servizio è formalmente specificato da un insieme di primitive (operazioni) che i processi utenti hanno a disposizione per accedere al servizio. Queste primitive istruiscono il servizio a eseguire alcune azioni o a riferire quelle prese da entità di pari livello.

Primitiva	Significato
LISTEN	Attesa bloccante di una connessione in arrivo
CONNECT	Stabilisce una connessione con un peer in attesa
ACCEPT	Accetta una richiesta di connessione da un peer
RECEIVE	Attesa bloccante per un messaggio in arrivo
SEND	Manda un messaggio al peer
DISCONNECT	Termina una connessione

Relazione tra servizi e protocolli

Servizi e protocolli sono concetti distinti.

Un servizio è un insieme di primitive (operazioni) che un livello offre a quello superiore. Il servizio definisce quali operazioni il livello sia in grado di offrire su richiesta dei suoi utenti, ma non dice nulla di come queste operazioni sono implementate.

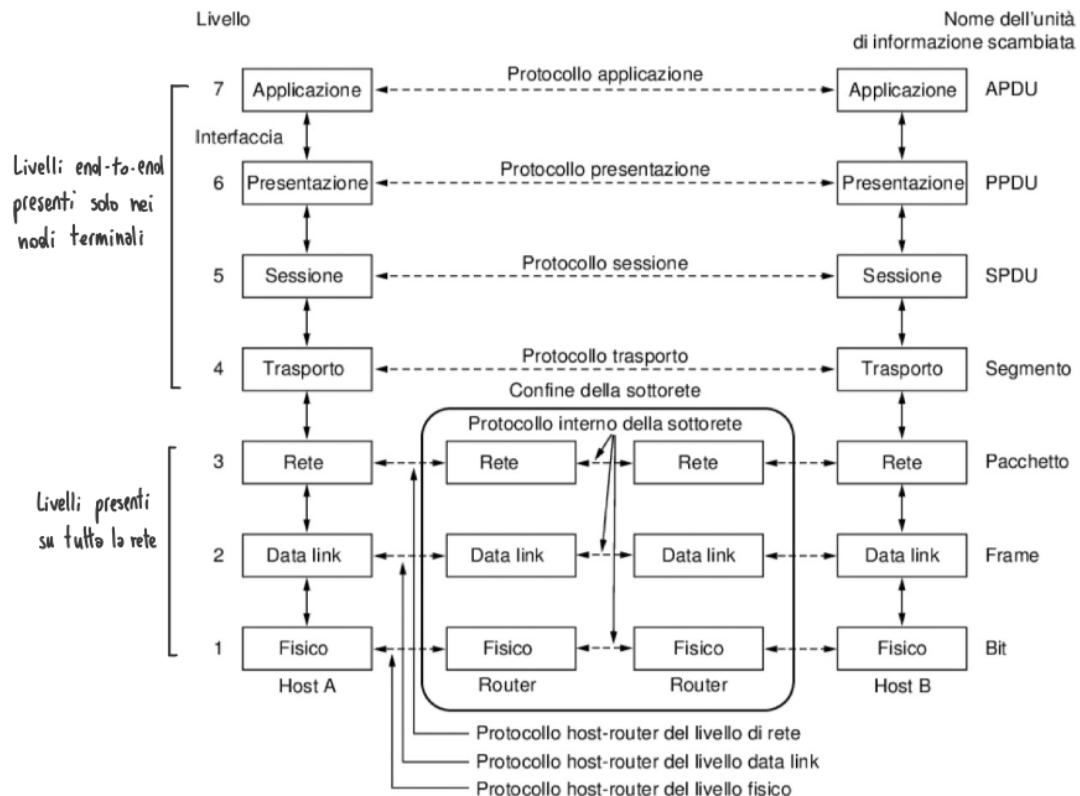
Un protocollo, invece, è un insieme di regole che controllano il formato e il significato dei pacchetti o messaggi scambiati tra le entità pari all'interno di un livello.

Le entità usano i protocolli per implementare le loro definizioni dei servizi. Sono libere di cambiare i loro protocolli, a patto di non cambiare il servizio visibile agli utenti. In questo modo, il servizio e il protocollo sono completamente disaccoppiati.

▼ 10.2 - Modelli di riferimento

Il modello di riferimento OSI

Il modello OSI (open system interconnection) si fonda su una proposta sviluppata dall'international standards organization (ISO) come primo passo verso la standardizzazione internazionale dei protocolli impiegati nei diversi livelli. Il modello OSI ha sette livelli.



1) Il livello fisico

Il livello fisico si occupa della trasmissione di bit grezzi sul canale di comunicazione. Le specifiche riguardano per lo più interfacce meccaniche o elettriche e temporizzazioni, oltre al mezzo di trasmissione che si trova sotto al livello fisico.

2) Il livello data link

Il compito principale dello strato data link consiste nel cercare di rilevare, per quanto possibile, gli errori di trasmissione così da evitare di trasmettere questi errori riconosciuti al livello superiore. L'obiettivo è raggiunto forzando il trasmettitore a suddividere i dati d'ingresso in data frame che vengono trasmessi sequenzialmente. Se il servizio è affidabile, il ricevitore conferma

la corretta ricezione di ciascun frame rimandando indietro un acknowledgment frame

3) Il livello di rete

Il livello di rete controlla il funzionamento della sottorete. Si occupa della modalità con cui i pacchetti sono inoltrati dalla sorgente alla destinazione.

Quando nella sottorete sono presenti contemporaneamente troppi pacchetti, creano delle congestioni: questo controllo spetta al livello di rete, per consentire la comunicazione tra reti eterogenee (routing).

4) Il livello di trasporto

La funzione essenziale del livello di trasporto è quella di accettare dati dal livello superiore, dividerli in unità più piccole quando necessario, passarle al livello di rete e assicurarsi che tutti i segmenti arrivino correttamente a destinazione. Il livello di trasporto copre tutto il percorso da sorgente a destinazione; inoltre stabilisce che tipo di servizio offre al livello sessione e, in definitiva, agli utenti della rete.

5) Il livello sessione

Il livello sessione permette a utenti su computer diversi di stabilire tra loro una sessione. Le sessioni offrono diversi servizi, tra cui: controllo del dialogo (tenere traccia di quando è il turno di trasmettere e quando di ricevere), gestione dei token (evitare che le due parti tentino la stessa operazione critica al medesimo istante) e sincronizzazione (supervisionare una lunga trasmissione per consentire la sua ripresa dal punto in cui si è interrotta a causa di un crash e del seguente recupero).

6) Il livello presentazione

Si occupa della sintassi e della semantica dell'informazione trasmessa. Il livello presentazione gestisce queste strutture dati astratte e consente lo scambio e la definizione di strutture dati di livello superiore (per esempio transazioni bancarie).

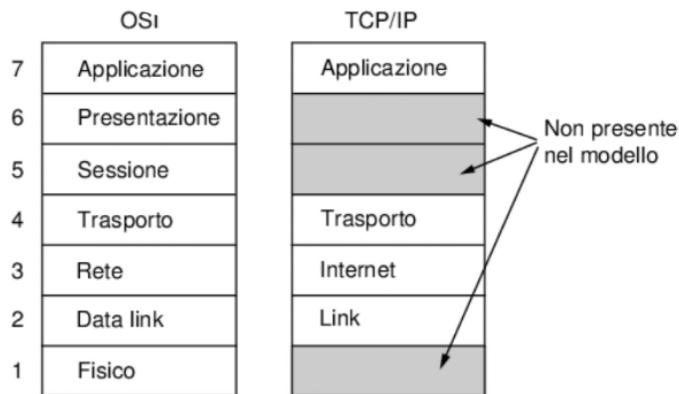
7) Il livello applicazione

Il livello applicazione comprende una varietà di protocolli comunemente richiesti dagli utenti. Un protocollo applicativo largamente usato è HTTP (hypertext transfer protocol), la base del World Wide Web.

Il modello di riferimento TCP/IP

Il modello di riferimento TCP/IP è il modello di riferimento del progenitore di tutte le reti di calcolatori geografiche, ARPANET, e il suo successore Internet. Poiché erano previste applicazioni con richieste divergenti, che

spaziavano dal trasferimento di file alla trasmissione della voce in tempo reale, era inoltre necessaria un'architettura flessibile.



1) Il livello link

Describe cosa devono fare i collegamenti (linee seriali e la classica Ethernet) per esaurire le necessità di questo livello internet senza connessione. È un interfaccia tra l'host e il mezzo trasmissivo.

2) Il livello internet

È il perno che tiene insieme l'intera architettura: il suo compito è permettere agli host di inviare pacchetti su qualsiasi rete e fare in modo che questi possano viaggiare indipendentemente verso la destinazione. Potrebbero persino arrivare con un ordine diverso da quello con cui sono stati spediti, e in questo caso è compito dei livelli superiori di riordinarli.

Il livello internet definisce un formato ufficiale per i pacchetti e un protocollo chiamato IP (internet control message protocol).

3) Il livello di trasporto

È progettato per consentire la comunicazione tra peer degli host sorgente e destinazione, come nel livello trasporto OSI.

In questo livello sono stati definiti due protocolli di trasporto end-to-end: il primo, TCP (transmission control protocol), è un protocollo affidabile orientato alla connessione che permette a un flusso di byte emessi da un computer di raggiungere senza errori qualsiasi altro computer sulla rete internet. Suddivide il flusso di byte entrante in messaggi e passa ciascun frammento al livello internet. A destinazione, il processo TCP ricevente ricompone il messaggio ricevuto per formare il flusso di uscita. TCP gestisce anche il controllo di flusso, per garantire che una sorgente veloce non possa congestionare un ricevente lento con una quantità di messaggi superiore a quelli che sa gestire. Il secondo protocollo di questo livello, UDP (user

datagram protocol), è un protocollo inaffidabile senza connessione per le applicazioni che non vogliono la garanzia di ordinamento e il controllo di flusso di TCP, ma preferiscono gestire queste funzioni in modo autonomo.

4) Il livello applicazione

Contiene tutti i protocolli di livello superiore. I primi gestivano un terminale virtuale (TELNET), lo scambio di file (FTP), la posta elettronica (SMTP). Con gli anni ne sono stati aggiunti molti altri, come il Domain Name System (DNS) che fa corrispondere i nomi degli host ai loro indirizzi di rete; HTTP, il protocollo per prelevare pagine sul WWW e RTP, il protocollo per trasmettere contenuti in tempo reale come audio e film.

Confronto tra i modelli OSI e TCP/IP

I modelli di riferimento OSI e TCP/IP hanno molto in comune. Sono entrambi basati sul concetto di pila di protocolli indipendenti, e la funzione dei livelli è grosso modo simile.

Nonostante queste somiglianze fondamentali, i due modelli hanno molte differenze. Nel modello OSI sono presenti tre concetti essenziali: servizi, interfacce, protocolli. Il modello TCP/IP in origine non faceva una netta distinzione tra questi tre concetti. La conseguenza è che nel modello OSI i protocolli sono nascosti meglio che nel modello TCP/IP e si possono sostituire con relativa facilità all'evolvere della tecnologia. Un'altra differenza consiste nella modalità di comunicazione: il modello OSI supporta nel livello di rete entrambi i tipi di comunicazione, orientata o meno alla connessione, ma nel livello di trasporto (importante perché visibile agli utenti) supporta solo la comunicazione orientata alla connessione. Il modello TCP/IP ha solo una modalità nel livello di rete (senza connessione), ma supporta entrambe in quello di trasporto offrendo così agli utenti una vera scelta.

Il modello ibrido

5	Application layer
4	Transport layer
3	Network layer
2	Data link layer
1	Physical layer

Questo modello ha cinque livelli che vanno dal livello fisico attraverso quelli di link, rete, trasporto fino al livello applicazione. Il livello fisico specifica

come trasmettere bit su differenti tipi di mezzi di trasporto in forma di segnali elettrici. Il livello link è coinvolto quando bisogna spedire messaggi di lunghezza finita direttamente tra computer connessi con un livello prefissato di affidabilità, Ethernet e 802.11 sono esempi di protocollo a livello link. Il livello di rete tratta come combinare link multipli nelle reti e nelle internetwork in modo da poter spedire pacchetti tra computer distanti. Questo include il compito di trovare un percorso attraverso cui mandare i pacchetti (protocollo IP). Il livello di trasporto si occupa di dare garanzie di consegna al livello di rete aumentandone l'affidabilità e fornisce astrazioni sulle modalità di consegna dei byte, che possono anche essere visti come una sequenza ordinata e affidabile (byte stream). (TCP). Infine, il livello applicazione contiene programmi che usano la rete.

Esempi di reti

Internet

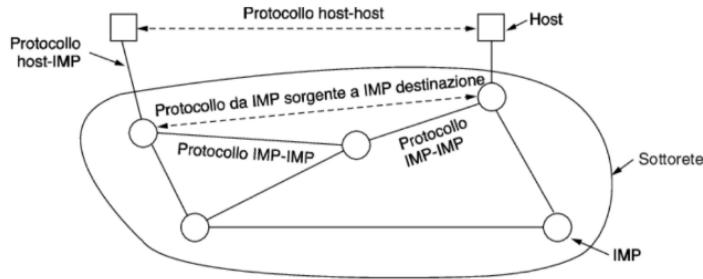
Internet non è una rete, ma una vasta raccolta di reti diverse che usano determinati protocolli e offrono certi servizi comuni. È un sistema inconsueto, che non ha un progettista e non è controllato da nessuno.

ARPANET

Rete di controllo che possa sopravvivere a una guerra nucleare.

Ai tempi le comunicazioni militari usavano la rete telefonica pubblica, considerata vulnerabile: la distruzione di poche centrali di alto livello avrebbe frammentato la rete in molte isole separate.

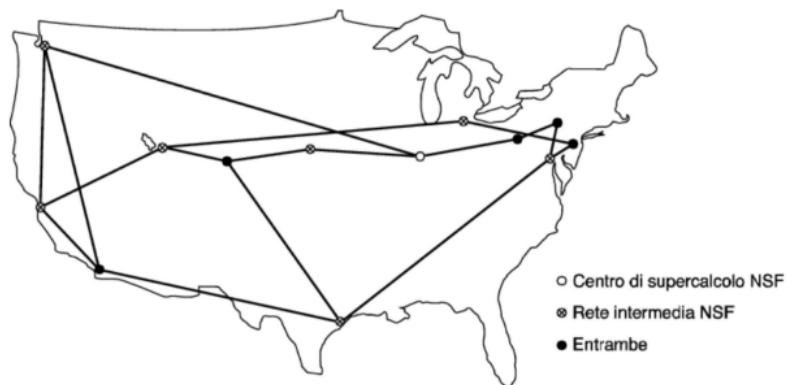
Nel 1960 venne proposto di usare una tecnologia digitale a commutazione di pacchetto. La sottorete sarebbe stata composta da minicomputer chiamati IMP (interface message processors) collegati da linee di trasmissione. Per consentire un'alta affidabilità ogni IMP doveva essere collegato ad almeno altri due IMP. Ogni nodo della rete doveva consistere in un IMP e un host, posti nella stessa stanza e collegati da un cavo. Ogni pacchetto sarebbe stato integralmente ricevuto prima dell'inoltro; quindi la sottorete fu la prima rete a commutazione di pacchetto del tipo store-and-forward. Il software della sottorete era composto dalla parte lato IMP della connessione tra IMP e host, dal protocollo IMP-IMP e da un protocollo da IMP sorgente a IMP destinazione progettato per migliorare l'affidabilità.



Al crescere delle dimensioni della rete rintracciare un host diventò sempre più difficoltoso, quindi fu creato il DNS (domain name system) per organizzare i computer in domini e abbinare i nomi degli host agli indirizzi IP.

NSFNET

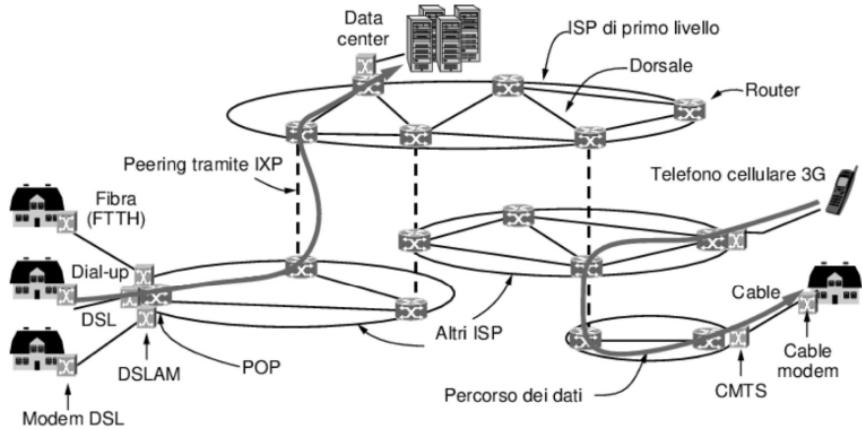
Negli ultimi anni '70 l'organismo statunitense NSF (national science foundation) si accorse dell'enorme impatto che ARPANET aveva sulla ricerca universitaria permettendo a scienziati di tutto il paese di condividere i dati e collaborare a progetti di ricerca. Negli ultimi anni '80 NSF progettò un successore di ARPANET che sarebbe stato aperto tutti i gruppi di ricerca universitari. A ogni supercomputer fu affiancato un fratellino, rappresentato da un microcomputer LSI-11 chiamato fuzzball. L'intera rete, composta da backbone e reti regionali, fu chiamata NSFNET. Il collegamento con ARPANET era realizzato da una connessione tra un IMP e un fuzzball situati nella sala machine della Carnegie Mellon University.



Architettura di Internet

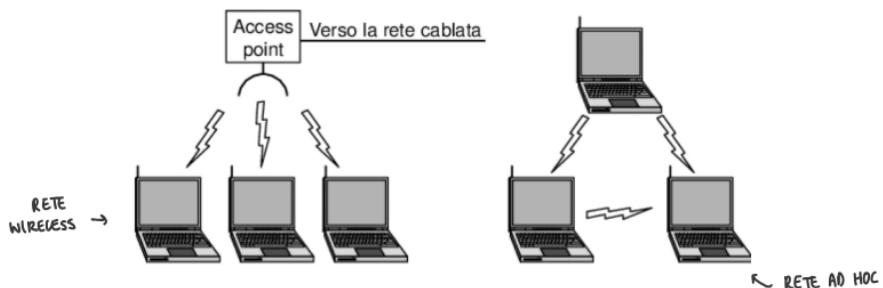
Per collegarsi a Internet, il computer si connette a un internet service provider, o più semplicemente ISP, che fornisce all'utente l'accesso a Internet o connettività. Questo permette al computer di scambiare pacchetti con tutti gli altri host accessibili in Internet.

Esistono molti tipi di accesso a Internet e di solito sono classificati in base alla banda che forniscono e al loro costo, anche se la caratteristica più importante è la connettività. La DSL, acronimo di digital subscriber line, riutilizza la linea telefonica di casa per la trasmissione di dati digitali. Il computer è connesso a un dispositivo chiamato modem DSL che effettua la conversione da pacchetti digitali a segnali analogici, che in questo modo passano sulla linea telefonica. Dall'altro capo, un dispositivo chiamato DSLAM (digital subscriber line access multiplexer) converte i segnali in pacchetti. La DSL è un modo di usare la linea telefonica locale che fornisce una banda più larga del dial-up, che consiste nell'inviare bit su una chiamata telefonica tradizionale, al posto di una comunicazione vocale. Il dial-up viene effettuato con molti tipi di modem ai due capi. La parola modem è l'abbreviazione di modulator demodulator e si riferisce a qualunque dispositivo che converta bit digitali in segnali analogici. Un altro metodo è quello di inviare segnali sul sistema della TV via cavo (cable). Il dispositivo collegato al capo situato in casa è chiamato cable modem, mentre il dispositivo all'altro capo (cable headend) è chiamato CMTS (cable modem termination system). Anche la modalità wireless viene utilizzata per l'accesso a Internet. Siamo ora in grado di mandare pacchetti da casa all'ISP. Chiamiamo il punto in cui i pacchetti entrano nella rete ISP con il nome di POP (point of presence) dell'ISP. Le reti ISP possono essere regionali, nazionali o internazionali; la loro architettura chiamata backbone (dorsale) dell'ISP. Se il pacchetto è destinato a un host servito direttamente dall'ISP, il pacchetto viene instradato sulla dorsale e consegnato all'host, altrimenti deve essere passato a un altro ISP. Gli ISP interconnettono le loro reti per scambiarsi traffico tramite gli IXP (Internet eXchange Point). Si dice che gli ISP fanno peering uno con l'altro. In cima alla "catena alimentare" c'è un pugno di aziende, quali AT&T e Sprint, che dispongono di grandi dorsali internazionali con migliaia di router connessi da fibre ottiche a banda larga. Questi ISP, chiamati tier 1 (di primo livello), non pagano il transito e formano una dorsale di Internet dato che chiunque altro deve collegarsi a loro per poter raggiungere il resto di Internet. Le aziende che forniscono molti contenuti come Google e Yahoo! concentrano i loro calcolatori in data center ben connessi al resto di Internet.



LAN wireless 802.11

Le reti 802.11 sono formate da client, come computer portatili e cellulari, e da infrastrutture chiamate AP (access point, punto di accesso) installate negli edifici. Tutti gli access point sono connessi a una rete cablata e tutte le comunicazioni tra i client passano attraverso almeno uno di loro. È anche possibile che due client che siano nel raggio radio uno dell'altro si parlino direttamente, come due computer in un ufficio, senza un access point. Questa configurazione è chiamata rete ad hoc (ad hoc network).



Standardizzazione delle reti

Esistono molti costruttori e fornitori di reti, ognuno con le proprie impostazioni. In assenza di coordinamento il caos sarebbe totale, e gli utenti non avrebbero nulla in mano. L'unica strada è raggiungere un accordo su alcuni standard di rete.

- Il Who's Who del mondo delle telecomunicazioni

ITU-T, il settore di standardizzazione delle telecomunicazioni, che si occupa di sistemi di telefonia e di scambio dati.

ITU-R, il settore delle radiocomunicazioni, si occupa di coordinare l'uso delle frequenze radio da parte di gruppi di interesse in competizione tra di loro ovunque nel mondo.

ITU-D, il settore dello sviluppo, promuove lo sviluppo di tecnologie di informazione e comunicazione terrestre per ridurre il "digital divide" tra le nazioni che hanno effettivamente accesso alle tecnologie dell'informazione e i paesi con accesso limitato.

- Il Who's Who del mondo degli standard internazionali

Gli standard internazionali sono definiti e pubblicati da ISO (international standards organization). ISO definisce standard su una vastissima gamma di argomenti. Per quanto riguarda le telecomunicazioni ISO e ITU-T spesso collaborano, per evitare di emettere ufficialmente due standard internazionali incompatibili tra loro.

Un altro ente importante nel mondo degli standard è l'IEEE (institute of electrical and electronics engineers). Il comitato IEEE 802 ha standardizzato molti tipi di LAN:

Number	Topic
802.1	Overview and architecture of LANs
802.2 ↓	Logical link control
802.3 *	Ethernet
802.4 ↓	Token bus (was briefly used in manufacturing plants)
802.5	Token ring (IBM's entry into the LAN world)
802.6 ↓	Dual queue dual bus (early metropolitan area network)
802.7 ↓	Technical advisory group on broadband technologies
802.8 ↑	Technical advisory group on fiber optic technologies
802.9 ↓	Isochronous LANs (for real-time applications)
802.10 ↓	Virtual LANs and security
802.11 *	Wireless LANs
802.12 ↓	Demand priority (Hewlett-Packard's AnyLAN)
802.13	Unlucky number. Nobody wanted it
802.14 ↓	Cable modems (defunct: an industry consortium got there first)
802.15 *	Personal area networks (Bluetooth)
802.16 *	Broadband wireless
802.17	Resilient packet ring

→ I gruppi di lavoro 802.1 più importanti sono contrassegnati con * quelli marciti da ↓ sono in quiescenza. Quelli con ↑ hanno rinunciato all'incarico o si sono sciolti.

- Il Who's Who del mondo degli standard di Internet

La Internet mondiale ha un proprio meccanismo di standardizzazione, molto diverso da ITUT e ISO. È divisa in gruppi di lavoro, che riguardano nuove applicazioni, informazioni degli utenti, integrazione OSI, instradamento e indirizzamento, sicurezza, gestione della rete e standard.

Per quanto riguarda gli standard Web, il World Wide Web consortium (W3C) sviluppa protocolli e linee guida che facilitano la crescita a lungo termine del Web.

▼ 10.3 - Livello 2: Data Link layer

In questo capitolo studieremo i principi dell'architettura dello strato numero 2: lo strato data link. Discuteremo gli algoritmi per ottenere una comunicazione affidabile ed efficiente fra due macchine adiacenti attraverso lo strato data link. Con il termine adiacenti indichiamo il fatto che le due macchine sono connesse da un canale di comunicazione che agisce concettualmente come un cavo. La proprietà essenziale di un canale per assimilarlo a un cavo è che i bit vengano instradati nell'esatto ordine in cui sono stati inviati.

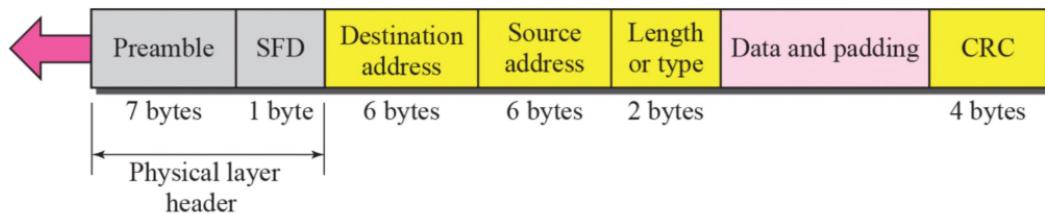
Questo strato è ulteriormente suddiviso in due sottostrati principali: il sottostrato MAC (Media Access Control) e il sottostrato LLC (Logical Link Control).

Token ring

Nel data link layer sono presenti nodi in cui vengono scambiati i token (token ring).

Preamble: 56 bits of alternating 1s and 0s.

SFD: Start frame delimiter, flag (10101011)



- Preamble: 8 byte, ognuno dei quali contiene la sequenza di bit 0101010, con eccezione dell'ultimo byte, in cui gli ultimi due bit sono impostati a 11; questo ultimo byte è chiamato delimitatore di inizio frame.
- Due indirizzi: uno per la destinazione e uno per la sorgente. Ognuno è lungo 6 byte.

I primi 3 byte del campo d'indirizzo sono usati per un OUI (organizationally unique identifier). I valori per questo campo sono assegnati da IEEE e indicano un produttore; il produttore assegna gli ultimi 3 byte dell'indirizzo.

d: Hexadecimal digit

d₁	d₂	:	d₃	d₄	:	d₅	d₆	:	d₇	d₈	:	d₉	d₁₀	:	d₁₁	d₁₂
----------------------	----------------------	---	----------------------	----------------------	---	----------------------	----------------------	---	----------------------	----------------------	---	----------------------	-----------------------	---	-----------------------	-----------------------

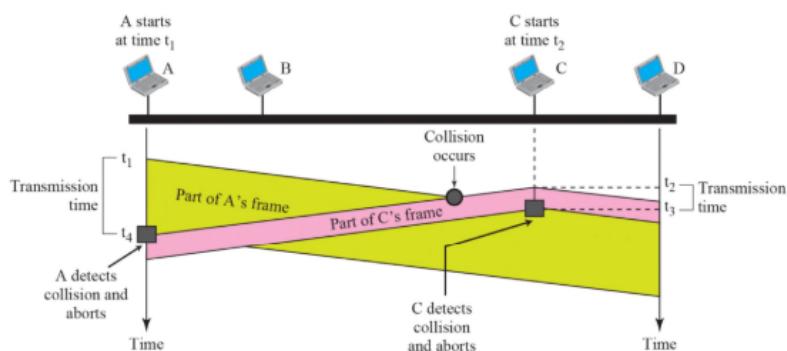
6 bytes = 12 hexadecimal digits = 48 bits

Nell'indirizzo di destinazione il primo bit è uno 0 per indirizzi ordinari e 1 per un gruppo di indirizzi. Quando un frame viene inviato a un indirizzo di gruppo, tutte le stazioni di quel gruppo lo ricevono. La trasmissione diretta a un gruppo di stazioni è definita multicast. L'indirizzo composto da tutti bit 1 è riservato per le trasmissioni broadcast.

- Type (tipo) o Length (lunghezza), a seconda che il frame sia Ethernet o IEEE 802.3. Ethernet usa un campo Type per indicare al ricevente cosa fare col frame: ad esempio il codice 0x0800 significa che i dati contengono un pacchetto IPv4.
- Data, lungo fino a 1500 byte.

Quando rileva una collisione, un transceiver tronca il frame corrente: ciò significa che sul cavo compaiono continuamente bit sparsi e pezzi di frame. Per aiutare a distinguere i frame validi dalla spazzatura, Ethernet richiede che i frame validi siano lunghi almeno 64 byte dall'indirizzo di destinazione al checksum inclusi.

Ethernet classica usa l'algoritmo CSMA/CD (Carrier Sensing Multiple Access - Collision Detection). Con ciò si intende che le stazioni controllano il canale quando hanno un frame da spedire e lo spediscono non appena il canale risulta libero; monitorano il canale per trovare collisioni mentre spediscono. Se c'è una collisione, interrompono la trasmissione con un breve messaggio caotico per poi ritrasmettere dopo un periodo di tempo casuale.



- Checksum: codice di rilevazione di errore che si usa per determinare se i bit del frame siano stati ricevuti correttamente.

Evoluzione di Ethernet

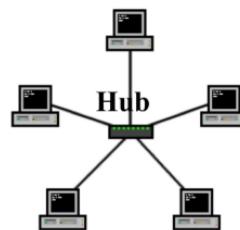
La prima rete LAN, formata da un singolo cavo coassiale, fu chiamata Ethernet. Aziende come Intel proposero uno standard di tale rete chiamato DIX, il quale poi con un piccolo cambiamento diventò lo standard IEEE 802.3.

Evoluzione della topologia di Ethernet

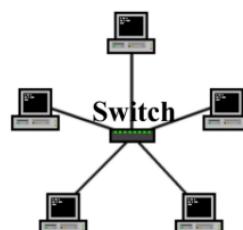
- Bus: cavo coassiale spesso con giunzioni a T e connettori BNC. Un singolo dominio di collisione, necessario il CSMA/CD



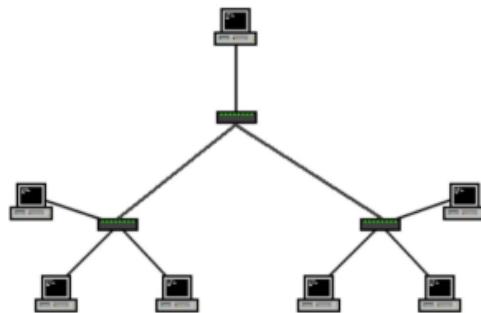
- Stella fisica: i nodi sono connessi ad un hub (ripetitore) al centro della stella. Un hub connette elettricamente tutti i cavi collegati come se fossero saldati assieme (è logicamente equivalente ad un bus). Un singolo dominio di collisione, necessario il CSMA/CD.



- Stella logica: i nodi sono connessi ad uno switch al centro della stella. A differenza di un hub, uno switch indirizza il frame verso la giusta porta di destinazione. Inoltre lo switch fa in modo che ogni nodo ha il suo dominio di collisione, dunque più nodi possono spedire contemporaneamente e CSMA/CD non è necessario.



- Strutture gerarchiche: ad esempio uno switch centrale che connette tre edifici, poi uno switch che connette tre piani, e ogni piano ha uno switch che serve tre aree.



Evoluzione di Ethernet

Ethernet classica si insinuava attraverso un edificio sotto forma di un singolo lungo cavo al quale tutti i computer erano collegati. Il primo tipo, comunemente chiamata thick Ethernet (Ethernet spessa), fu seguita da thin Ethernet (Ethernet sottile). Ogni variante di Ethernet ha una lunghezza massima del cavo per ogni segmento (cioè la lunghezza su cui non si effettua una amplificazione) su cui il segnale si propaga. Per permettere di costruire reti più grandi, più cavi possono essere connessi attraverso repeater. Su ognuno di quei cavi le informazioni sono spedite usando la codifica Manchester.

I numerosi problemi della struttura a lungo cavo portarono all'utilizzo di reti a stella fisica e logica, le quali differenze sono state spiegate in precedenza.

Dopo l'ethernet classica (10 Mbit/s) si è passati poi alla fast ethernet (100 Mbit/s). L'idea alla base era di mantenere tutti i vecchi formati di frame per calcoli, interfacce e regole procedurali, riducendo semplicemente il tempo bit da 100 ns a 10 ns. I tipi di cavi che vennero supportati furono i doppini di categoria 3 e 5. Sono utilizzati solo due doppini per stazione: uno da e uno verso l'hub, dunque il sistema può essere full-duplex (si può trasmettere e ricevere contemporaneamente). Non sono utilizzate ne una codifica binaria diretta né la codifica Manchester, viene invece utilizzata la codifica 4B/5B: 4 bit di dati sono codificati con 5 bit di segnale e spediti a 125 MHz per ottenere 100 Mbps.

Lo standard gigabit Ethernet fu ratificato da IEEE nel 1999. Gli obiettivi del comitato per gigabit Ethernet erano gli stessi posti per fast Ethernet: aumentare le prestazioni di dieci volte mantenendo la compatibilità con tutti gli standard Ethernet esistenti. Anche qui si può usare sia l'half duplex (con l'hub, serve CSMA/CD) che il full duplex (con lo switch, non serve CSMA/CD). Poiché un frame può essere trasmesso 100 volte più velocemente che nella Ethernet classica, la lunghezza massima del cavo risulta 100 volte più corta: 25 metri. Questa restrizione di lunghezza fu così dolorosa da richiedere l'aggiunta di due funzionalità allo standard, al fine di

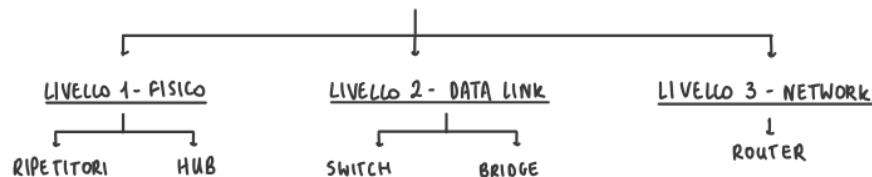
aumentare la lunghezza massima del cavo a 200 metri: La prima funzionalità, chiamata carrier extension, indica essenzialmente all'hardware di aggiungere dei dati di riempimento dopo il frame normale, in modo da estendere la dimensione del pacchetto fino a 512 byte. La seconda funzionalità, chiamata frame bursting, permette a un trasmettente di inviare una sequenza concatenata di più frame in una singola trasmissione. Per inviare messaggi a una velocità di 1 Gbps vengono utilizzati cavi in rame brevi e schermati e fibre ottiche.

Lo standard 10 gigabit Ethernet fu ratificata da IEEE per la prima volta nel 2002. Tutte le versioni di 10-gigabit Ethernet supportano solo la modalità full-duplex. Tutte le versioni di 10GbE inviano un flusso sequenziale di dati prodotto mischiando i bit e poi applicando una codifica 64B/66B.

Ethernet 40/100 Gbit/s, ratificata nel 2010, raggiunge velocità decisamente più elevate, supportando sempre solo operazioni full duplex, preservando tutte le altre caratteristiche.

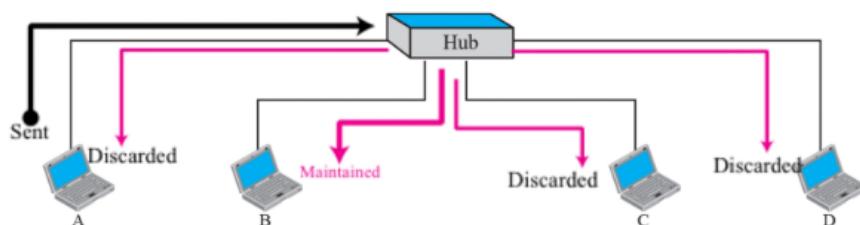
Dispositivi di commutazione

I dispositivi di interconnessione si classificano in base al livello in cui operano.



Hub

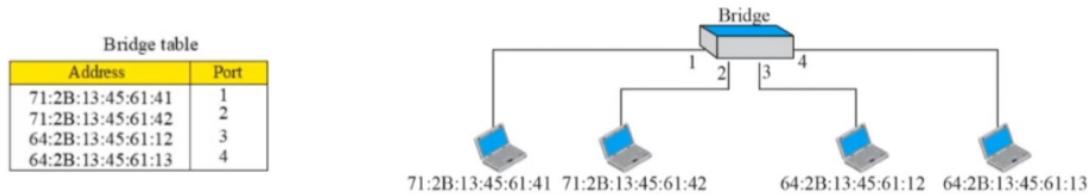
Un hub, o ripetitore, non tratta frame o pacchetti, ma solo bit. Inoltre ogni bit a tutte le porte, e non ha capacità di filtraggio.



Switch

Gli switch Ethernet sono un modo attuale di chiamare i bridge. I bridge operano a livello 2, quindi esaminano gli indirizzi del livello data link per

inoltrare i frame (non i pacchetti). Con un bridge si possono unire due LAN separate. Il bridge possiede una tabella e usa gli indirizzi MAC per trovare la porta a cui inoltrare il frame di input.

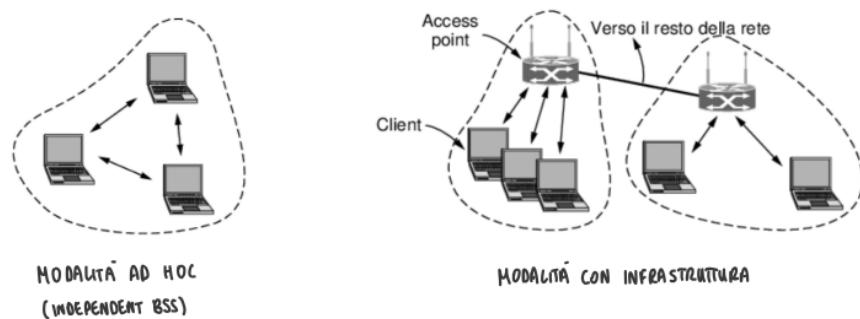


Nel learning bridge si parte da una tabella vuota, che si riempie automaticamente man mano che "conosce" gli indirizzi delle varie porte a cui invia/riceve frame.

IEEE 802.11

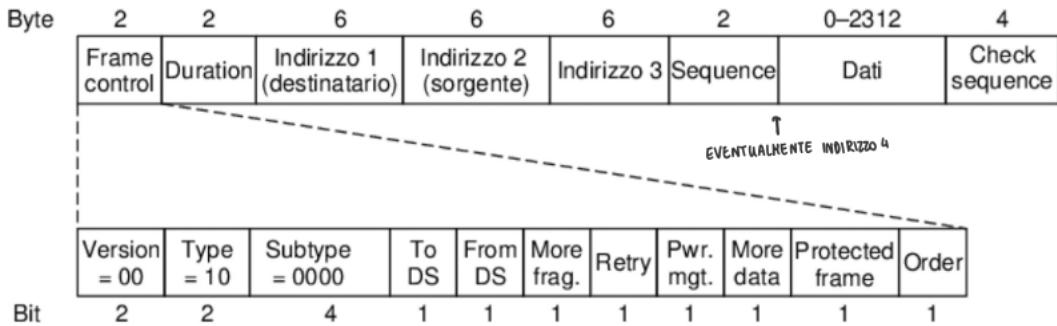
Il principale standard per le LAN Wireless è IEEE 802.11, meglio conosciuto come WiFi.

Le reti 802.11 possono essere utilizzate in due modalità: nella modalità con infrastruttura ogni client è associato a un AP (access point) connesso a sua volta all'altra rete. Il client spedisce e riceve i propri pacchetti attraverso l'AP. L'altra modalità prende il nome di rete ad hoc, Questa modalità consiste in una collezione di computer associati tra loro che possono spedirsi direttamente i frame. Dal momento che l'accesso a Internet è la killer application per le reti wireless, le reti ad hoc non sono molto popolari.



Struttura del frame di 802.11

Lo standard 802.11 definisce tre diverse classi di frame: dati, controllo e gestione. Ognuna ha un'intestazione composta da una varietà di campi utilizzati all'interno del sottolivello MAC.



- Frame control, costituito da 11 sottocampi: Protocol version impostato a 00, Type (con possibili valori dati, controllo e gestione) e Subtype, To DS e From DS sono fissati a indicare se il frame stia andando o venendo dalla rete connessa all'AP, More fragments significa che seguiranno più frammenti, Retry indica la ritrasmissione di un frame spedito in precedenza, Power management indica che il mittente sta andando in modalità power-save, More data indica che il mittente ha altri dati per il ricevente, Protected Frame indica che il corpo del frame è stato criptato per sicurezza, Order indica al ricevente che il livello superiore si aspetta che la sequenza di frame arrivi rigorosamente in ordine.
- Duration, indica per quanto tempo il frame e il suo acknowledgment occuperanno il canale, misurato in microsecondi.
- 3 indirizzi. Il primo indirizzo è il ricevente e il secondo è il mittente. Se il frame viene inviato ad un AP viene utilizzato anche il terzo indirizzo per indicare al punto finale di destinazione. Vi è un eventuale quarto indirizzo.
- Sequence numera i frame in modo da poter individuare duplicati.
- Dati contiene il payload.
- Frame check sequence, che è lo stesso CRC a 32 bit di 802.3.

I frame di gestione hanno lo stesso formato dei frame di dati, più un formato per la porzione dati che varia a seconda di Subtype.

I frame di controllo sono brevi. Come tutti i frame hanno i campi Frame control, Duration, e Frame check sequence. Potrebbero non avere nessuno spazio per i dati, in quanto la maggior parte delle informazioni chiave viene trasmessa nel campo Subtype (ACK, RTS e CTS).

Nel caso di canali condivisi, sono possibili due approcci:

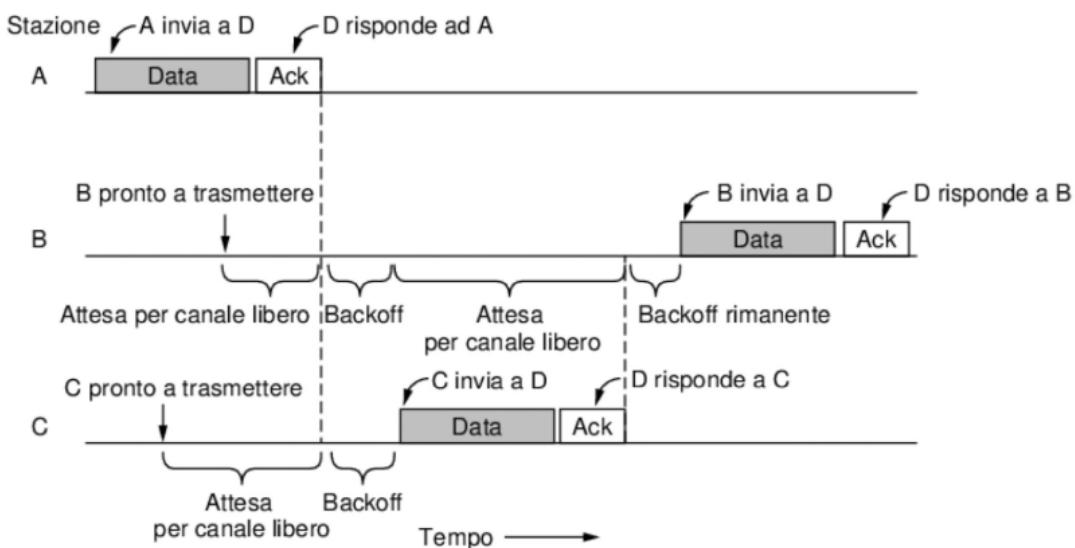
- Polling: l'AP "chiede" alle stazioni se hanno frame da inviare.

- Contention: le stazioni devono competere per usare il medium, si utilizza l'algoritmo CSMA-CA (CSMA Collision Avoidance).

Protocollo del sottolivello MAC di 802.11

Con Ethernet, una stazione attende finché il mezzo trasmissivo non diventa libero e inizia a trasmettere. Se non riceve alcun picco di rumore mentre spedisce i primi 64 byte, il frame è quasi certamente stato consegnato correttamente.

Con il wireless questo meccanismo di rilevamento delle collisioni non funziona, dunque 802.11 prova a evitare le collisioni con un protocollo chiamato CSMA/CA (CSMA with collision avoidance). In questo protocollo la stazione attende finché il canale è libero (senza segnale) per un certo periodo e poi inizia a contare alla rovescia gli slot di backoff (numero scelto casualmente tra 0 e 15). Se durante il conto alla rovescia vengono inviati dei frame, la stazione mette in pausa il conteggio.



Nell'esempio in figura notiamo che la stazione A è la prima a spedire un frame. Mentre A spedisce, le stazioni B e C diventano pronte per spedire. Vedono che il canale è occupato e attendono che si liberi. Poco dopo che A riceve un acknowledgement, il canale diventa libero. Tuttavia, invece di spedire subito un frame e farlo collidere, entrambe B e C eseguono un random backoff. C sceglie un backoff breve, perciò spedisce per primo. B mette in pausa il conto alla rovescia quando rileva che C sta utilizzando il canale e ricomincia quando C ha ricevuto un acknowledgement. B completa presto il backoff e spedisce il suo frame.

Questo modo di operare è chiamato DCF (distributed coordination function) perché ogni stazione agisce indipendentemente, senza alcun controllo.

centrale. Lo standard include anche un modo di operare opzionale chiamato PCF (point coordination function) in cui l'AP controlla tutta l'attività nella sua cella.

Un problema di questo tipo di trasmissione è che i raggi di trasmissione (portata radio) di ogni stazione possono essere differenti, per questo è possibile incorrere nel problema del terminale nascosto: le trasmissioni che avvengono in una parte di una cella potrebbero non essere ricevute in qualche altro punto nella stessa cella. Per questo motivo una stazione può non rilevare una trasmissione e pensare di poter trasmettere verso un'altra stazione generando una collisione.



Per evitare il problema del terminale nascosto 802.11 definisce la rilevazione del canale come composta sia dalla rilevazione fisica sia da quella virtuale. La rilevazione fisica semplicemente controlla il mezzo trasmisivo per capire se c'è un segnale valido. Ogni frame possiede un campo NAV (network allocation vector) che indica quanto tempo sia necessario per completare la sequenza di cui il frame fa parte. Le stazioni che vedono il frame sapranno che il canale resterà occupato per il periodo di tempo indicato dal NAV, indipendentemente dal fatto che sia rilevato un segnale fisico.

Siccome i canali wireless sono molto rumorosi, e la probabilità che un frame venga trasferito senza bit errati diminuisce con l'aumentare della dimensione del frame, 802.11 ammette la frammentazione dei frame, i quali vengono suddivisi in parti più piccole, ognuna dotata del proprio checksum e acknowledgement individuale, il quale consente di ritrasmettere solo i frammenti danneggiati.

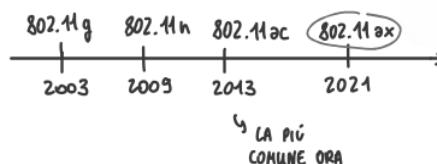
PCF e DCF possono coesistere dentro la stessa cella: dopo la trasmissione di un frame è richiesta una pausa prima di qualsiasi nuova trasmissione; lo standard prevede quattro intervalli ognuno dei quali ha una funzione specifica. L'intervallo più breve è chiamato SIFS (Short InterFrame Spacing)

ed è utilizzato per concordare i turni tra le parti coinvolte nella singola conversazione. C'è sempre un'unica stazione autorizzata a rispondere dopo un intervallo SIFS. Se questa possibilità non viene sfruttata e l'intervallo PIFS (PCF InterFrame Spacing) finisce, la stazione base può inviare un frame di segnalazione o di interrogazione. Se la stazione base non ha nulla da dire e trascorre un intervallo DIFS (DCF InterFrame Spacing), qualunque stazione può tentare di acquisire il controllo del canale per inviare un nuovo frame. In questo caso si applicano le solite regole della contesa e in caso di collisioni può essere attivato l'algoritmo di backoff esponenziale binario. L'ultimo intervallo, EIFS (Extended InterFrame Spacing), è utilizzato solo da una stazione che ha appena ricevuto un frame danneggiato o sconosciuto, e serve per annunciarlo.

Un altro problema possibile in 802.11 è quello del terminale esposto: B vuole inviare dati a C, quindi ascolta il canale; quando sente una trasmissione, conclude erroneamente di non poterlo fare anche se A potrebbe in effetti trasmettere a D (non visibile). Questa decisione spreca un'opportunità di trasmissione.



Sommario dell'evoluzione di 802.11



▼ 10.4 - Livello 3: Network layer

Introduzione al livello 3: network layer

Il livello rete si occupa del trasporto dei pacchetti lungo tutto il percorso dall'origine alla destinazione finale. Questa funzione è chiaramente distinta da quella del livello data link, che ha il solo obiettivo di spostare i frame da un capo all'altro di un cavo. Per raggiungere i propri obiettivi, il livello di rete deve conoscere la topologia della rete di comunicazione (l'insieme dei router e dei collegamenti) e scegliere i percorsi appropriati attraverso di essa, evitando di sovraccaricare alcune linee di comunicazione e lasciando altre completamente libere.

Per capire come fa, è necessario rivedere alcuni concetti già visti in precedenza:

- **Switching:** quando un messaggio raggiunge un dispositivo di connessione, bisogna decidere da quale porta di output il pacchetto deve essere inviato. In poche parole, il dispositivo si comporta come uno switch che connette una porta ad un'altra porta.

Esistono più tipologie di switching:

- **Circuit switching** (commutazione a circuito): il messaggio è mandato da sorgente a destinazione senza essere diviso in pacchetti.
- **Packet switching** (commutazione a pacchetto): bit o byte organizzati in pacchetti che vengono poi riassemblati una volta raggiunta la destinazione.

Sono possibili due diverse tipologie di organizzazioni, che dipendono dal tipo di servizio offerto. Se il servizio è senza connessione, i pacchetti (chiamati datagram) sono inoltrati nella rete individualmente e instradati indipendentemente l'uno dall'altro, senza scegliere un percorso predefinito. In questo caso ogni router ha una tabella interna che indica dove devono essere inviati i pacchetti diretti a ogni possibile destinazione. Questa tabella si aggiorna con il variare di alcune condizioni (es. ingorgo in un certo percorso).

Se il servizio è orientato alla connessione, prima di inviare i pacchetti si deve stabilire un percorso che colleghi il router sorgente al router destinazione, chiamato circuito virtuale. Questo percorso è utilizzato per tutto il traffico che scorre attraverso la connessione, ma non è proprietario di un solo circuito virtuale, ma può essere utilizzato da più circuiti virtuali. Ogni pacchetto contiene al suo interno un identificatore del circuito virtuale utilizzato, e ogni router contiene nella propria tabella di routing il prossimo router a cui inviare un pacchetto facente parte di uno specifico circuito virtuale.

Indirizzi

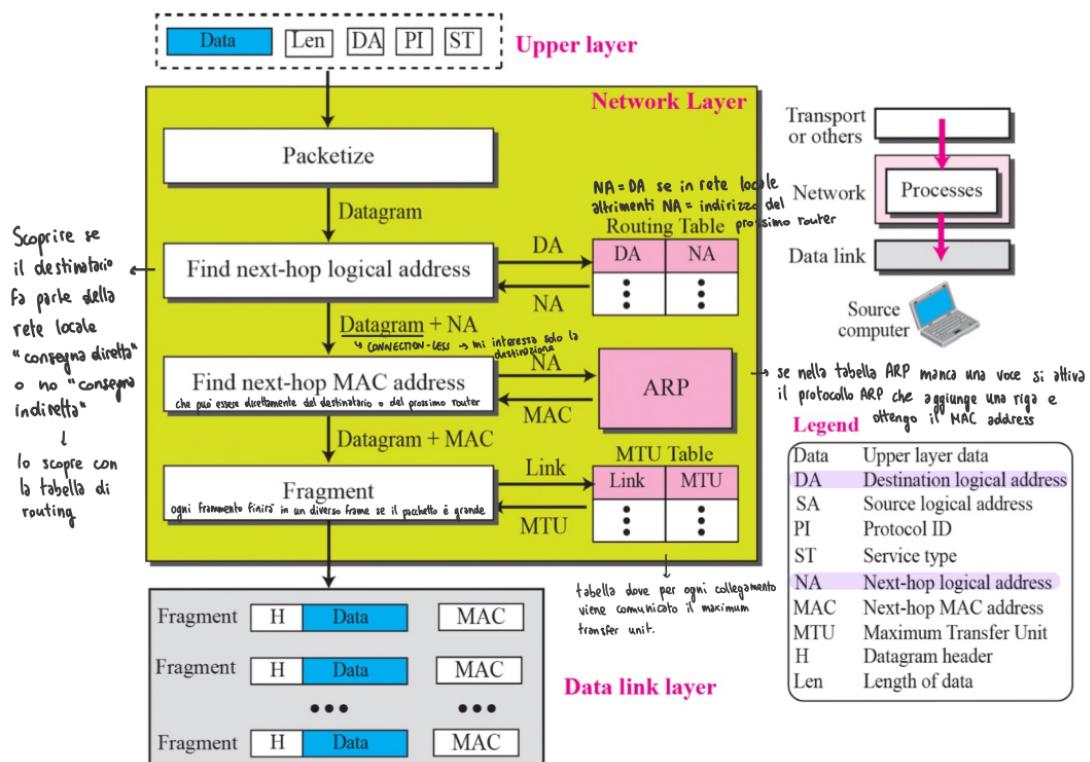
Esistono diversi tipi di indirizzi a seconda del livello:

- Network layer → indirizzi logici (IP). Possono essere di tipo IPv4 e IPv6, i quali non sono compatibili tra loro ma ogni scheda di rete li ha entrambi. Sono indirizzi dinamici, che vengono associati per un periodo limitato. Ogni indirizzo IP è unico e viene dunque associato ad un indirizzo MAC specifico, per cui esiste la tabella ARP, presente in ogni computer, la quale associa ad ogni indirizzo IP il suo indirizzo MAC.
- Data link layer → indirizzi fisici (MAC). Sono indirizzi statici, come un numero di serie.

Gli indirizzi possono inoltre essere:

- Pubblici: visibili su Internet, univoci in tutto il mondo, assegnati da qualcuno.
- Privati: non visibili su Internet, visibilità limitata alla rete locale.

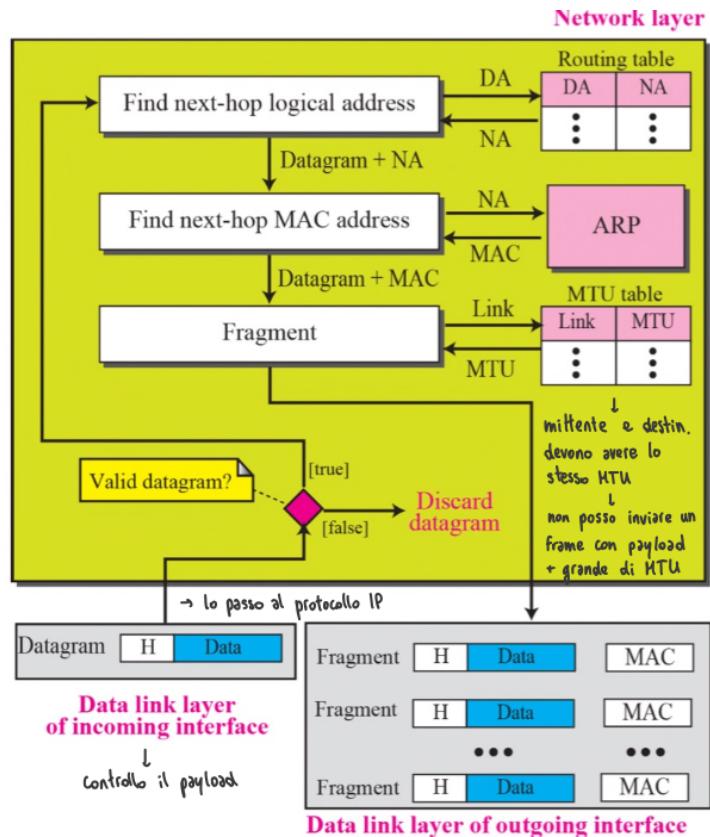
Funzionamento illustrato dei protocolli a livello 3



Dalla figura notiamo che i dati dell'applicazione vengono incapsulati in un datagram. Il sistema verifica poi se la destinazione è nella rete locale, e in tal caso l'indirizzo di destinazione viene utilizzato direttamente, altrimenti si utilizza l'indirizzo del prossimo router ricavato interrogando la tabella di

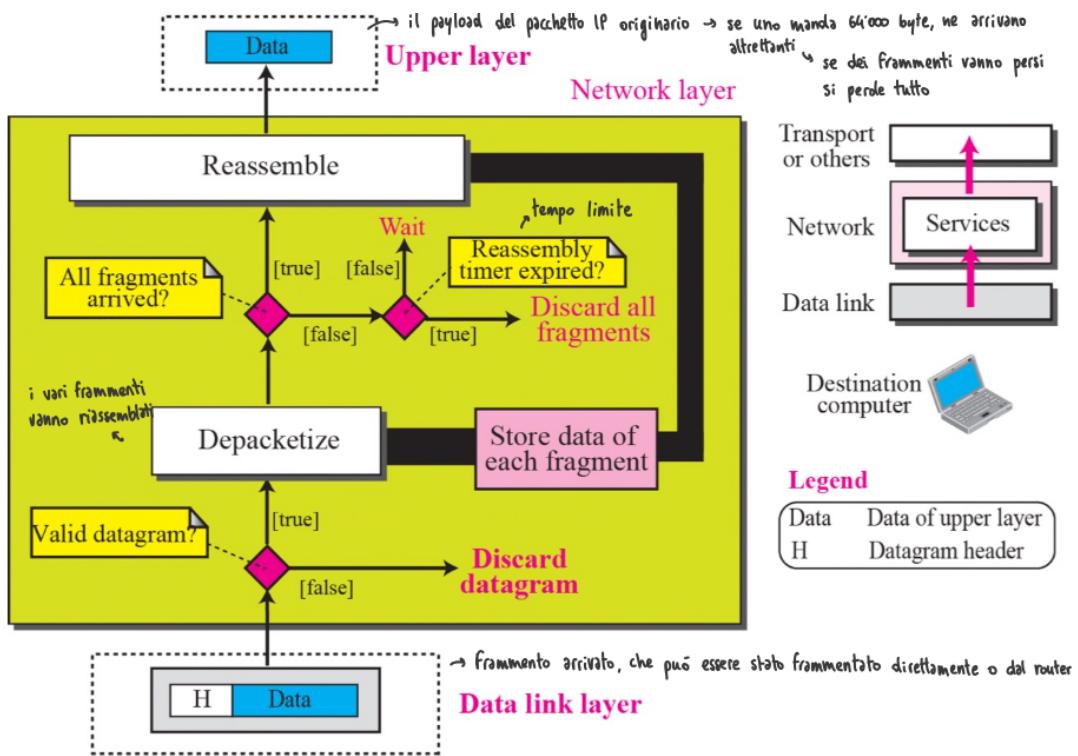
routing. Successivamente, utilizzando la tabella ARP si trova l'indirizzo MAC associato all'indirizzo logico del prossimo router. Si controlla dunque se i datagrammi sono troppo grandi per l'MTU (Maximum Transfer Unit), e in tal caso vengono frammentati, per poi, infine, essere trasmessi al livello data link.

Il router opera dunque nel seguente modo:



Una volta arrivato il datagramma dall'interfaccia di input data link, viene controllata la sua validità, e in caso positivo tramite la routing e ARP table vengono calcolati gli indirizzi logico e fisico del prossimo hop. Infine viene impacchettato il tutto in un frammento che viene inviato all'interfaccia di output del livello data link.

Una volta arrivato il frammento nel computer di destinazione questo è il processo adottato per elaborarlo:



Il frammento arriva dal livello data link, e dopo aver controllato che questo sia valido viene depacchettizzato e controllato che tutti i frammenti della comunicazione siano arrivati. In caso positivo si riassembra il payload del pacchetto originario, altrimenti si aspettano che arrivino altri frammenti.

Protocollo IPv4

Indirizzi IPv4

L'identificatore usato nel livello IP del protocollo TCP/IP per identificare ogni dispositivo connesso ad Internet si chiama indirizzo IP, il quale è un indirizzo a 32-bit che definisce universalmente e unicamente la connessione di un host o router ad Internet, infatti un dispositivo può avere tanti indirizzi IP quante connessioni (solitamente un host ne ha uno, ma ad esempio un router ne ha tanti).

Ogni indirizzo a 32 bit è composto di una parte di rete di lunghezza variabile nei primi bit e di una parte per l'host negli ultimi. La parte di rete ha lo stesso valore per tutti gli host di una singola rete, e tale blocco di bit che la identifica è chiamato prefisso. Per convenzione viene scritta dopo il prefisso dell'indirizzo IP con uno slash / seguito dalla lunghezza in bit della parte dedicata alla rete. La lunghezza del prefisso corrisponde a una maschera binaria di 1 nella parte di rete (subnet mask, maschera di sottorete).

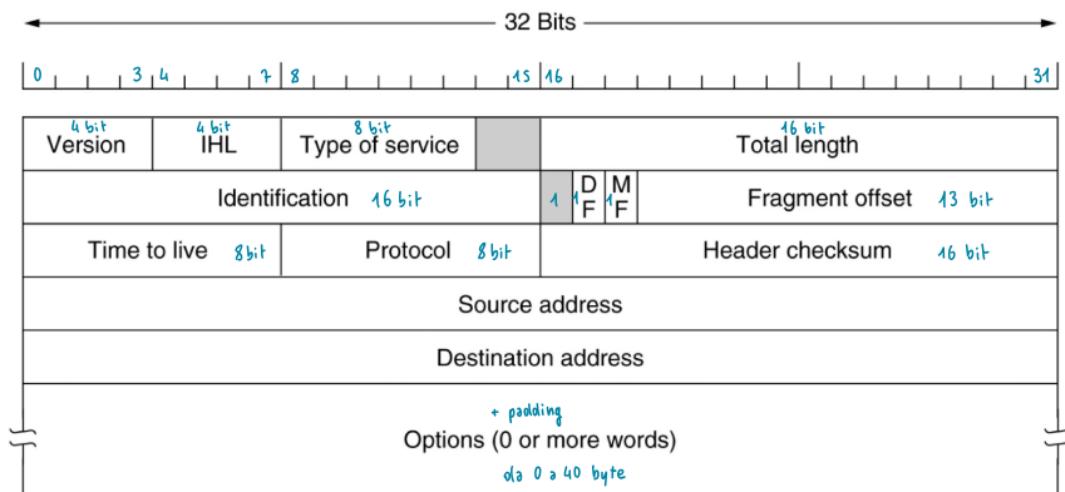
Ci sono ulteriori indirizzi speciali: l'indirizzo IP 0.0.0.0 è utilizzato dagli host al momento del boot e significa "questa rete" o "questo host". L'indirizzo

composto da tutti 1, o 255.255.255.255, viene utilizzato per indicare tutti gli host di una rete. Infine tutti gli indirizzi espressi nella forma 127 sono riservati per effettuare controlli all'interno dell'host stesso (loopback).

Vengono inoltre riservati alcuni indirizzi IP come indirizzi privati, e nessun pacchetto contenente questi indirizzi può apparire su internet. Questi intervalli di indirizzi sono i seguenti: 10.0.0.0 - 10.255.255.255/8, 172.16.0.0 - 172.31.255.255/12, 192.168.0.0 - 192.168.255.255.

Datagramma IPv4

Un datagramma IPv4 è costituito da una parte di intestazione e da un corpo (il payload). L'intestazione ha una parte fissa di 20 byte e una parte opzionale di lunghezza variabile:



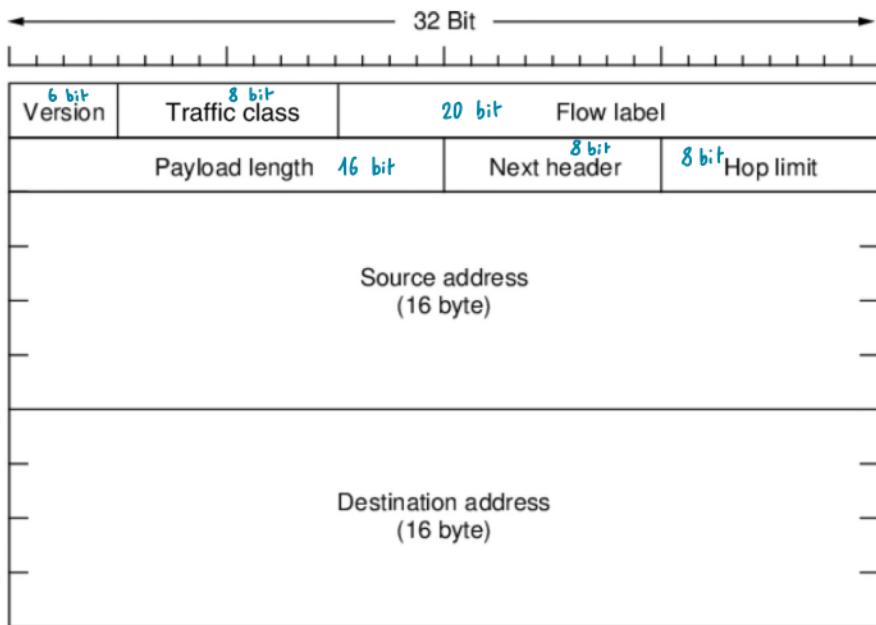
Il campo IHL indica la lunghezza dell'intestazione, Identification (identificazione) serve all'host di destinazione per determinare a quale datagramma appartiene il frammento appena arrivato. DF "don't fragment" rappresenta un ordine che impone ai router di non dividere in frammenti il datagramma; MF "more fragments" è impostato a 1 per tutti i frammenti tranne l'ultimo, per comunicare quando sono arrivati tutti i frammenti. Fragment offset indica la posizione del frammento nel datagramma corrente. Il campo Time to live è un contatore utilizzato per limitare la vita di un pacchetto. Il campo Protocol indica quale processo di trasporto è in attesa di quei dati. Il campo Options (opzioni) contiene opzioni di lunghezza variabile. Ognuna inizia con un codice di 1 byte che la identifica.

Protocollo IPv6

Datagramma IPv4

Molto presto gli indirizzi a disposizione di IPv4 iniziarono a scarseggiare, fino a terminare. IPv6 è un progetto che usa indirizzi a 128 bit.

L'intestazione di un pacchetto IPv6 è la seguente:



Il campo Traffic class è usato per distinguere le classi di servizio dei pacchetti con differenti richieste di consegna in tempo reale. Il campo Flow label (etichetta di flusso) consente a una sorgente e a una destinazione di marcare un gruppo di pacchetti che, avendo gli stessi requisiti, devono essere trattati allo stesso modo. Il motivo per cui è stato possibile semplificare l'intestazione è che ci possono essere altre intestazioni estese (opzionali). Il campo Next header (intestazione successiva) indica quale delle sei (per il momento) intestazioni estese, se presente, segue l'intestazione corrente. Il campo Hop limit (limite di hop) è utilizzato per impedire ai pacchetti di vivere per sempre.

Indirizzi IPv6

Per scrivere gli indirizzi a 16 byte è stata scelta una nuova notazione. Gli indirizzi sono scritti come otto gruppi di quattro cifre esadecimale separate da due punti:

8000 : 0000 : 0000 : 0000 : 0123 : 4567 : 89AB : CDEF

NAT: Network Address Translation

La soluzione a lungo termine dell'esaurimento degli indirizzi IP è che tutta Internet passi a IPv6. Questa transizione sta procedendo lentamente e ci

vorranno anni prima che si completi. Di conseguenza era necessario trovare una soluzione rapida attuabile in tempi brevi: si chiama NAT (network address translation). L'idea di base di NAT è assegnare a ogni azienda o casa un singolo indirizzo IP (o, al massimo, un piccolo numero di indirizzi) per traffico di Internet. Dentro la rete del cliente, ogni computer riceve un indirizzo IP unico (utilizza i 3 intervalli di indirizzi privati), utilizzato per instradare il traffico interno alla rete locale. Tuttavia quando un pacchetto sta per lasciare la rete locale per dirigersi verso l'ISP viene eseguita una traduzione di indirizzo dall'unico indirizzo IP interno a quello pubblico condiviso.

La traduzione avviene tramite l'utilizzo di apparati NAT, i quali convertono gli indirizzi interni in indirizzi esterni.

Siccome la maggior parte dei pacchetti appartengono a connessioni di tipo TCP/UDP, i quali protocolli prevedono una porta sorgente e una destinazione, il NAT, oltre a cambiare l'indirizzo IP sorgente interno con un IP pubblico, deve cambiare anche la porta sorgente del pacchetto con la porta sorgente originale. La porta sorgente viene cambiata perché due dispositivi diversi nella rete interna potrebbero usare la stessa porta (es. 5000). Quando un pacchetto arriva dall'esterno, il NAT usa la porta sorgente modificata come indice per trovare l'IP interno e la porta originale nella sua tabella. Infine, il pacchetto viene inviato al dispositivo interno corrispondente.

NAT TABLE → (ESEMPIO)	Original Source	Original Port	Translated Source	Translated Source Port	Destination	Destination Port
			IP of NAT device	It works as a label to identify the original pair (IP, port)	Unaltered	Unaltered
192.168.0.1 (PRIVATO)	1400 ←	137.204.57.12 LA RISPOSTA ARRIVA QUI	137.204.57.12	2405 "ETICHETTA"	173.194.40.81 (PUBBLICO)	80 (PUBBLICO)
192.168.0.2	1653	137.204.57.12	137.204.57.12	2406 RIESCE A INSTRADARE ANCHE CON STESSO INDIRIZZO	173.194.40.81	80 PORTA DEL SERVER WEB
192.168.0.5	1653	137.204.57.12	137.204.57.12	2407 ↓ PORTE ESSENZIALI!	173.194.40.81	80

▼ 10.5 - Livello 4: Transport layer

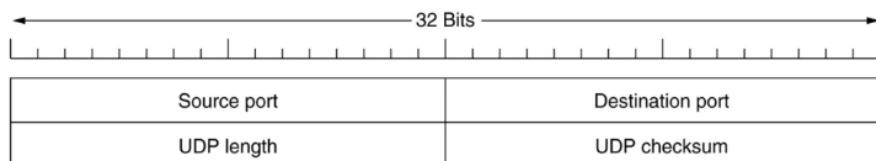
Il livello di trasporto si basa sul livello di rete per fornire il trasporto dei dati da un processo su una macchina sorgente a un processo su una macchina di destinazione con un livello di affidabilità desiderato e indipendente dalle reti fisiche attualmente in uso.

Il livello di trasporto Internet possiede due protocolli principali: un protocollo non orientato alla connessione e uno orientato alla connessione. Il protocollo non orientato alla connessione è UDP; non fa praticamente altro che spedire pacchetti tra le applicazioni, permettendo loro di costruirsi sopra i propri protocolli come necessario. Il protocollo orientato alla connessione è TCP. Fa praticamente tutto: crea connessioni e aumenta l'affidabilità con le ritrasmissioni, implementa anche il controllo di fusso e il controllo di congestione, tutto per conto delle applicazioni che lo usano.

UDP (User Datagram Protocol)

UDP trasmette segmenti costituiti da un'intestazione di 8 byte seguita dal payload.

L'intestazione è composta da due porte che servono per identificare gli endpoint all'interno dei computer di sorgente e destinazione. I campi delle porte sorgente e destinazione sono il vantaggio principale dell'utilizzo di UDP rispetto a IP, perché consentono la consegna del segmento all'interno del frame alla corretta applicazione. La porta sorgente serve principalmente quando si deve inviare una risposta al mittente.



Intestazione UDP.

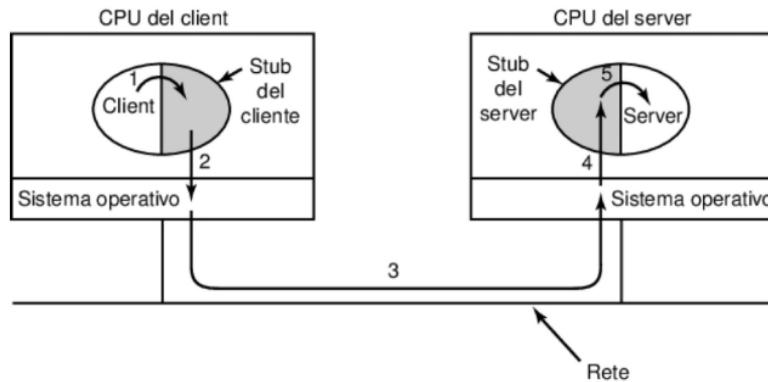
Il protocollo UDP non si occupa del controllo di flusso, del controllo di congestione o della ritrasmissione dopo la ricezione di un segmento errato; questi compiti sono lasciati ai processi utente.

RPC (Remote Procedure Call)

Quando un processo sulla macchina 1 chiama una procedura sulla macchina 2, il processo chiamante su 1 viene sospeso e l'esecuzione della procedura chiamata avviene su 2. Questa tecnica è nota come RPC (remote procedure call, chiamata a procedura remota) ed è diventata la base per molte applicazioni di rete. Tradizionalmente la procedura chiamante è nota come client, mentre la procedura chiamata è nota come server; L'idea alla base di RPC è eseguire una chiamata a procedura remota in maniera più simile possibile a una locale.

Nella forma più semplice, per chiamare una procedura remota, il programma client deve essere associato a una piccola procedura di libreria, chiamata

client stub, che rappresenta la procedura del server nello spazio di indirizzamento del client. allo stesso modo il server è associato a una procedura chiamata server stub. Queste procedure nascondono il fatto che la chiamata a procedura dal client al server non sia locale.



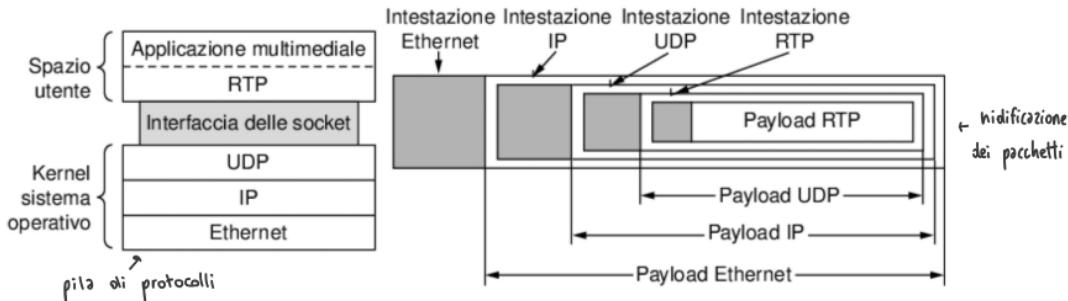
L'esecuzione di una RPC è la seguente: il client effettua una chiamata a procedura locale al suo stub. Il client stub inserisce i parametri in un messaggio ed effettua una chiamata di sistema al fine di far inviare il messaggio dal kernel della macchina client a quello della macchine server. Una volta arrivato il pacchetto in ingresso viene inviato al server stub, il quale effettua una chiamata a procedura locale per inviare il pacchetto alla destinazione finale all'interno del server.

RTP (Real Time Transport Protocol)

RTP è un protocollo di trasporto in tempo reale per applicazioni di vario genere.

RTP viene normalmente eseguito nello spazio utente appoggiandosi su UDP. Un'applicazione multimediale è composta di più flussi audio, video, di testo e forse di altro tipo. Questi flussi vengono passati alla libreria RTP, la quale esegue il multiplexing dei fusti e li codifica in pacchetti RTP, che vengono poi inseriti in una socket. Nella parte di kernel che si occupa della socket vengono generati pacchetti UDP che encapsulano quelli RTP, che vengono poi passati al livello IP per essere trasmessi. Il processo inverso ha luogo al ricevente e l'applicazione multimediale riceve i dati dalla libreria RTP e si occupa di rappresentarli.

RTP può essere descritto come un protocollo di trasporto implementato a livello applicazione.



RTCP (Real Time Transport Control Protocol)

RTP si accompagna a un protocollo chiamato RTCP, che gestisce le retroazioni verso la sorgente, la sincronizzazione e l'interfaccia utente. Non trasporta campioni multimediali. Tra le diverse funzioni, viene utilizzato per fornire alla sorgente di flusso un feedback su ritardi, larghezza di banda, congestione e altre proprietà di rete. Fornendo un feedback continuo, gli algoritmi di codifica si possono adattare costantemente per fornire la migliore qualità.

TCP (Transmission Control Protocol)

TCP è stato progettato appositamente per fornire un flusso di byte affidabile end-to-end su una internetwork affidabile. I dati sono visti come un flusso di byte, anche se sono organizzati in pacchetti chiamati segmenti. Il TCP permette una consegna ordinata dei dati, con ritrasmissione dei dati persi. Inoltre prevede un controllo di flusso, ovvero non vengono spediti più dati di quanti ne può ricevere la sorgente in un dato momento, e un controllo di congestione, ovvero si cerca di non spedire più dati di quanti ne può smaltire la rete.

Il servizio TCP è ottenuto con la creazione di punti terminali di un sistema di comunicazione da parte di mittente e ricevente, chiamati socket. Ogni socket possiede un indirizzo composto dall'indirizzo IP dell'host e da una porta di 16 bit locale all'host. Per ottenere il servizio TCP si deve stabilire esplicitamente una connessione tra una socket su una macchina e una socket su un'altra macchina. Una socket può essere usata per più connessioni contemporaneamente. I numeri di porta minori di 1024 sono riservati per servizi standard, e prendono il nome di well-known port. Tutte le connessioni TCP sono di tipo full-duplex punto a punto. Full-duplex indica che il traffico può procedere in entrambe le direzioni contemporaneamente. Punto a punto significa che ogni connessione ha esattamente due punti terminali, dunque il TCP non supporta il multicast o il broadcast.

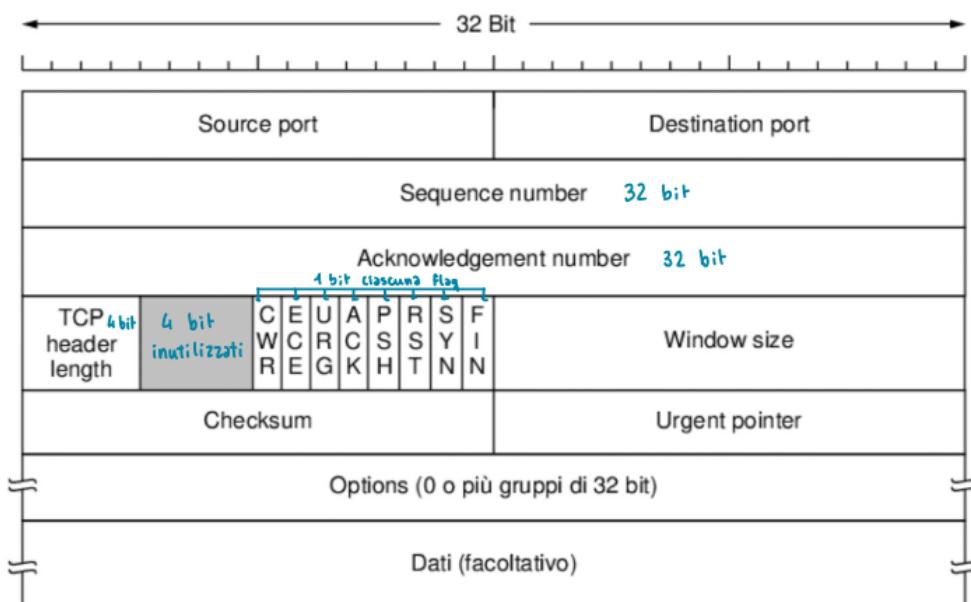
Affidabilità del TCP

L'affidabilità del TCP è data dal seguente funzionamento: ogni segmento ha un numero di sequenza che indica la posizione del primo byte del segmento all'interno del flusso di dati. Questo numero aiuta il ricevitore a ordinare correttamente i segmenti ricevuti. Quando il ricevitore riceve un segmento, invia una conferma (ACK) al mittente, la quale contiene il numero del prossimo byte che il ricevitore si aspetta di ricevere. Questo numero indica implicitamente che tutti i byte precedenti sono stati ricevuti correttamente (es. se il ricevitore invia un ACK per il byte 101, significa che ha ricevuto correttamente tutti i byte fino al 100). Se un segmento viene perso durante la trasmissione, il ricevitore invia degli ACK duplicati (DupACK) per segnalare al mittente che non ha ricevuto il segmento atteso (es. se il ricevitore si aspetta il byte 101 ma riceve segmenti successivi, continuerà a inviare ACK per il byte 101 finché non riceve il segmento mancante).

Il RTT è il tempo che intercorre tra l'invio di un segmento da parte del mittente e la ricezione dell'ACK corrispondente. Un RTT più breve significa che la connessione è più veloce ed efficiente. In alcune situazioni, il ricevitore non invia un ACK per ogni segmento ricevuto, ma aspetta di ricevere due segmenti prima di inviare un ACK. Questo si chiama delayed ACK e viene utilizzato per ridurre il numero di pacchetti ACK inviati, migliorando l'efficienza della rete.

Il protocollo TCP

Un segmento TCP consiste di un'intestazione fissa di 20 byte seguita da zero o più byte di dati.

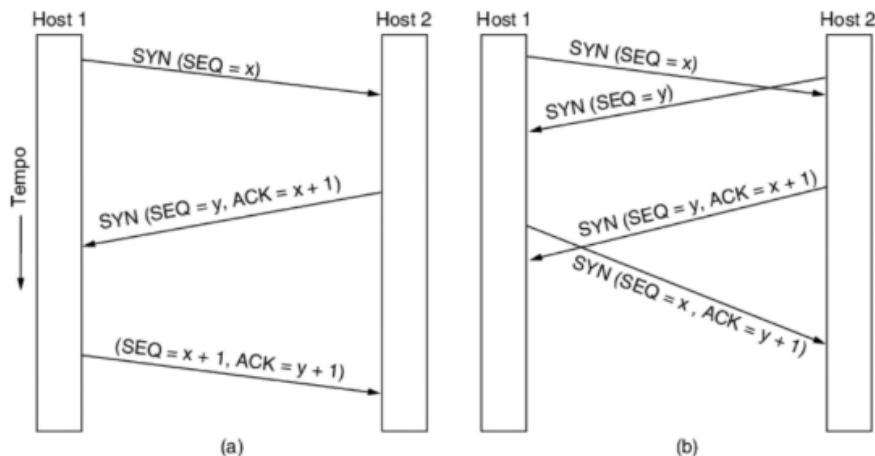


Intestazione TCP.

Tra gli 8 bit di flag troviamo ACK, per indicare che il segmento contiene un acknowledgement, SYN, utilizzato per stabilire una connessione, e FIN, utilizzato per rilasciare una connessione.

Instaurazione della connessione TCP

Per stabilire una connessione, un lato attende in modo passivo una connessione in ingresso eseguendo nell'ordine le primitive LISTEN e ACCEPT. L'altro lato segue una primitiva CONNECT, specificando l'indirizzo IP e la porta a cui vuole connettersi. La primitiva CONNECT invia un segmento TCP con il bit SYN a 1 e il bit ACK a 0, poi attende una risposta. Quando questo segmento arriva a destinazione, l'entità TCP controlla se esiste un processo che ha eseguito una LISTEN sulla porta indicata nel campo Destination port e in caso negativo invia una risposta con il bit RST a 1 per rifiutare la connessione. Se invece un processo è in ascolto sulla porta, gli viene dato il segmento TCP in ingresso; quindi può accettare o rifiutare la connessione. Se accetta, viene restituito al mittente un segmento con il bit SYN a 1 e il bit ACK a 1.



Nel caso in cui due host tentino contemporaneamente di stabilire una connessione tra le stesse due socket, il risultato di questi eventi è la costituzione di una sola connessione, non due, perché le connessioni sono identificate dai loro punti terminali.

Stati della connessione TCP

I passi richiesti per stabilire e rilasciare le connessioni possono essere rappresentati in una macchina a stati finiti con gli 11 stati elencati:

Stato	Descrizione
CLOSED	Nessuna connessione è attiva o in sospeso
LISTEN	Il server è in attesa di una chiamata in ingresso
SYN RCVD	È arrivata una richiesta di connessione; in attesa di ACK
SYN SENT	L'applicazione ha iniziato ad aprire una connessione
ESTABLISHED	Il normale stato di trasferimento dei dati
FIN WAIT 1	L'applicazione ha detto di aver terminato
FIN WAIT 2	L'altro lato ha accettato il rilascio
TIME WAIT	Attende la scadenza di tutti i pacchetti
CLOSING	Entrambi i lati hanno cercato di chiudere contemporaneamente
CLOSE WAIT	L'altro lato ha iniziato il rilascio
LAST ACK	Attende la scadenza di tutti i pacchetti

Controllo di flusso

Per il controllo di flusso il protocollo TCP utilizza un protocollo a finestre scorrevoli. Ciò che il mittente può spedire dipende non solo dagli ack ricevuti ma anche dallo spazio disponibile nel buffer del ricevente, indicato dal campo window size che la stazione di destinazione manda al mittente in ogni segmento in modo che il mittente possa regolare la trasmissione dei segmenti.

Quando la dimensione della finestra è 0 il mittente non può spedire dati; può comunque spedire un segmento di un byte per fare in modo che l'altra stazione ripeta qual è il prossimo byte atteso e segnali la dimensione della finestra; questo segmento sonda (che prende il nome di window probe) viene mandato quando scade un apposito timer (fatto partire quando il mittente trasmette un segmento) per evitare problemi nel caso fosse stata segnalata una finestra di dimensione 0 e poi fosse andato perduto il segmento con l'aggiornamento della dimensione della finestra.

W è il numero massimo di segmenti che possono essere spediti dopo l'ultimo confermato. Se $W = 1$ la velocità di trasmissione è data da $T_x = W \cdot RTT$, mentre se $W > 1$ la velocità di trasmissione è $T_x = \frac{W}{RTT}$. Se si vuole che la velocità di trasmissione corrisponda alla banda disponibile, occorre utilizzare una $W = B \cdot RTT$, dove B è la banda disponibile. W è inoltre determinato come il minimo tra due valori, cwnd, ovvero la dimensione della finestra di congestione della rete, e rwnd, ossia la dimensione della finestra di ricezione. Otteniamo dunque che $T_x = \min(\frac{cwnd}{RTT}, \frac{rwnd}{RTT})$.

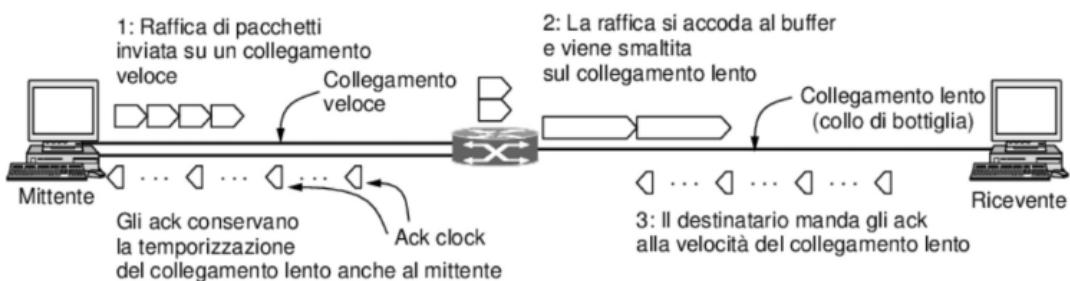
Controllo di congestione

Quando il carico applicato a qualsiasi rete è più di quello che questa riesce a gestire, si crea

una congestione. Il livello di rete rileva la congestione quando le code sui router aumentano e cerca di gestirla scartando dei pacchetti. È compito del livello di trasporto ricevere i segnali di congestione dal livello di rete e rallentare il tasso del traffico inviato sulla rete.

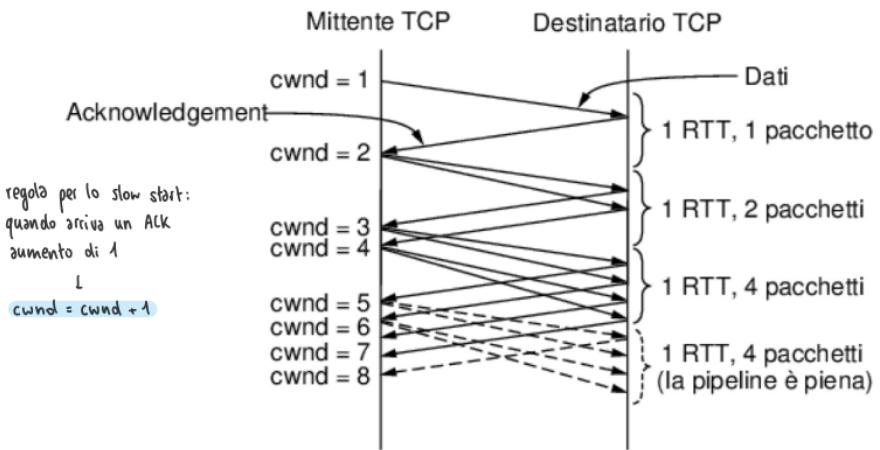
Il controllo della congestione di TCP utilizza una finestra in cui inserire i pacchetti che vengono scartati. Tale finestra ha una dimensione pari al numero di byte che il mittente può avere sulla rete in qualsiasi momento. Si ricordi che la finestra di congestione viene mantenuta in aggiunta alla finestra di controllo del flusso che specifica il numero di byte che il destinatario può inserire nel suo buffer, dunque il numero di byte che possono essere spediti è il più piccolo delle due finestre.

Un metodo adottato dal protocollo TCP per ridurre la congestione è quello dell'ack clock. Prendiamo per esempio un mittente che spedisce pacchetti su una rete veloce ad un destinatario su una rete lenta. Con l'inviare dei pacchetti questi vengono accodati nel buffer del router, in quanto la rete del destinatario è più lenta di quella del mittente. Una volta che i pacchetti raggiungono il destinatario, questo invia degli acknowledgement al mittente, e la velocità di trasmissione di questi ack viene mantenuta pari a quella di trasferimento nel collegamento lento. In questo modo, una volta che gli ack raggiungono il mittente, questo calcola il tasso a cui i pacchetti possono essere spediti sul collegamento più lento tramite la velocità di arrivo degli ack, e in questo modo invia i prossimi pacchetti nella rete rispettando tale tasso ed evitando congestioni.

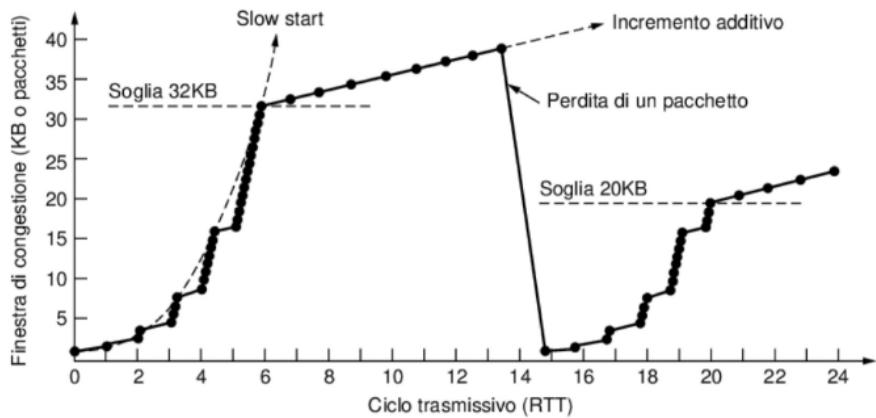


Un metodo per portare gradualmente il protocollo a regime è quello che utilizza l'algoritmo slow start. Quando viene stabilita una connessione, il mittente inizializza la finestra di congestione a un valore iniziale basso. Il mittente, quindi, spedisce la finestra iniziale. I pacchetti impiegheranno un round-trip time per ottenere i relativi acknowledgement. Per ogni segmento che ha ricevuto l'acknowledgement prima dello scadere del timer di ritrasmissione il mittente aggiunge alla finestra di congestione il valore in

byte di un segmento, dunque per ogni acknowledgement possono essere inviati altri due segmenti. La finestra di congestione raddoppia a ogni round-trip time. Lo slow start funziona bene su una gamma di velocità e round-trip time e usa un ack clock per allineare il tasso delle trasmissioni del mittente al percorso di rete.



Quando la rete si satura, i pacchetti possono iniziare a perdere. Quando il mittente non riceve una conferma per un pacchetto entro un certo tempo, entra in "timeout". Questo indica che c'è stata una congestione nella rete. Per evitare di sovraccaricare la rete, il mittente tiene traccia di una soglia chiamata "slow start threshold". Se il numero di pacchetti inviati supera questa soglia, il TCP cambia strategia e inizia a incrementare la finestra di congestione in modo più lento, aggiungendo solo un pacchetto per ogni ciclo di conferma, anziché raddoppiarla. Quando viene rilevata una perdita di pacchetti (per esempio, attraverso un timeout), la soglia di slow start viene ridotta a metà del valore della finestra di congestione attuale.



Questa tecnica che è stata appena descritta è chiamata TCP Tahoe, ma esistono altre tecniche di controllo della congestione. Un esempio è quello di TCP Reno, la quale si comporta come TCP Tahoe per la fase iniziale ma, alla perdita di un pacchetto, invece di reimpostare la finestra di congestione a 1 e ripartire con lo Slow Start utilizza il fast retrasmisit ritrasmettendo rapidamente il pacchetto mancante senza aspettare il timeout e riduce a metà la finestra di congestione e la aumenta con crescita lineare (fast recovery).

