

ELETTRONICA I

Prof. Antonio Pio Catalano – A.A. 2023/24

Luca Maria Incarnato

INDICE DEGLI ARGOMENTI

ELETTRONICA ANALOGICA

1. IL SEGNALE ELETTRICO (p. 3)
2. LA FISICA DEI SEMICONDUTTORI (p. 4)
3. LE CORRENTI ELETTRICHE NEI SEMICONDUTTORI (p. 8)
4. LA GIUNZIONE PN (p. 12)
5. IL DIODO (p. 17)
6. LA REGIONE DI BREAKDOWN E IL DIODO ZENER (p. 25)
7. IL TRANSISTORE BIPOLARE A GIUNZIONE (p. 28)
8. LE CARATTERISTICHE E LE REGIONI DI FUNZIONAMENTO DI UN BJT (p. 31)
9. RAPPRESENTAZIONE AD EMETTITORE COMUNE E L'EFFETTO EARLY (p. 35)
10. POLARIZZAZIONE DI UN BJT E CIRCUITI AMPLIFICATORI (p. 37)
11. IL MODELLO EQUIVALENTE A PICCOLO SEGNALE (p. 45)
12. L'AMPLIFICATORE AD EMETTITORE COMUNE A BJT (p. 48)
13. AMPLIFICATORI A COLLETTORE E A BASE COMUNE A BJT (p. 50)
14. LA RISPOSTA IN FREQUENZA DEGLI AMPLIFICATORI A BJT (p. 53)
15. IL MOSFET (p. 57)
16. LE CARATTERISTICHE TENSIONE – CORRENTE DI UN MOSFET (p. 60)
17. IL MOSFET A CANALE P E A SVUOTAMENTO (p. 65)
18. MODELLO A PICCOLO SEGNALE DEL MOSFET (p. 67)
19. AMPLIFICATORI ELEMENTARI A MOSFET (p. 71)
20. L'ELETTRONICA INTEGRATA E L'AMPLIFICATORE DIFFERENZIALE (p. 72)
21. L'AMPLIFICATORE OPERAZIONALE (p. 76)
22. COMPORTAMENTO IN FREQUENZA DELL'OPERAZIONALE (p. 89)
23. LA NON IDEALITÀ DELL'AMPLIFICATORE OPERAZIONALE (p. 91)
24. L'INTEGRATORE REALE E LA RETE DI REAZIONE GENERALIZZATA (p. 93)

ELETTRONICA DIGITALE

25. IL CONVERTITORE FLASH (p. 96)
26. INVERTITORI LOGICI (p. 96)
27. IL MOSFET COME BIPOLO DI CARICO (p. 103)
28. NAND E NOR IN LOGICA A RAPPORTO E L' INVERTITORE CMOS (p. 110)
29. PORTE E FUNZIONI LOGICHE COMPLESSE (p. 118)
30. CIRCUITI COMBINATORI: DI DECODIFICA E INSTRADAMENTO (p. 124)
31. CIRCUITI SEQUENZIALI: BISTABILI, LATCH E FLIP – FLOP (p. 129)
32. CIRCUITI SEQUENZIALI: MEMORIE NON VOLATILI (p. 139)

ELETTRONICA ANALOGICA

IL SEGNALE ELETTRICO

Il mondo dell'informazione poggia sul concetto di **segnaletico**, inteso come una **forma di tensione o corrente che**, manipolata secondo particolari schemi, **trasmette un'informazione**. Una delle possibili operazioni che si possono fare su un segnale è quella di **amplificazione**; tuttavia, se **il segnale è descritto da una tensione, il circuito che permette di amplificare questo segnale non è realizzabile da una rete passiva lineare**, dal momento in cui **non permette la restituzione di una tensione maggiore di quella fornita in ingresso**. Si può pensare di utilizzare un generatore controllato per eseguire un'operazione di amplificazione e a quel punto serve conoscere la natura delle grandezze fisiche di ingresso e di uscita per poter modulare al meglio la costante di guadagno:

$$y(t) = Cx(t)$$

$$y \approx v \wedge x \approx v \Rightarrow C = \left[\frac{V}{V} \right]$$

$$y \approx i \wedge x \approx i \Rightarrow C = \left[\frac{A}{A} \right]$$

$$y \approx v \wedge x \approx i \Rightarrow C = \left[\frac{V}{A} = \Omega \right]$$

$$y \approx i \wedge x \approx v \Rightarrow C = \left[\frac{A}{V} = S \right]$$

Nei primi due casi si parlerà di guadagno di tensione e guadagno di corrente, mentre negli ultimi due di guadagno di transresistenza e guadagno di transconduttanza.

L'amplificazione di una tensione (dualmente di una corrente) **amplifica anche l'energia del sistema?** Ovviamente no, altrimenti ci sarebbe un guadagno di energia senza alcun dispendio, andando contro il principio di conservazione dell'energia. Nelle condizioni attuali potrebbe essere fatta la seguente valutazione:

$$\begin{cases} P_i(t) = v_i(t)i_i(t) \\ P_o(t) = v_o(t)i_o(t) = A v_i(t)i_i(t) \end{cases} \Rightarrow P_o(t) > P_i(t)$$

E sarebbe un **affermazione errata**. Questa disuguaglianza potrebbe essere **alimentata anche da eventuali dissipazioni per effetto Joule**, di potenza $P_d(t)$; ci si chiede, allora, cosa possa fornire al circuito una potenza $P_a(t)$ che equilibri il bilancio energetico:

$$P_i(t) + P_a(t) = P_d(t) + P_o(t)$$

La potenza in questione viene definita **potenza di alimentazione** e può essere fornita da **batterie o da particolari circuiti detti alimentatori stabilizzati**, che trasformano la tensione sinusoidale della rete in una tensione (o corrente) continua.

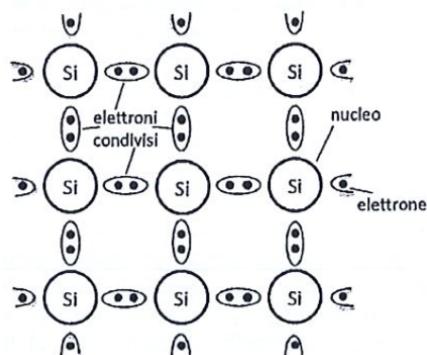
Da tutto questo ragionamento va estratto il concetto per cui **un qualsiasi circuito elettronico, analogico o digitale, non può funzionare se non è correttamente alimentato**; per il momento ci si disinteressa del tipo di alimentazione e del motivo per cui conviene stabilizzare la tensione (o corrente) in ingresso.

Prima di concludere, si vuole fare una necessaria **distinzione tra circuiti** (più generalmente **sistemi**) **lineari e non lineari**; i primi sono circuiti in cui **tensioni e correnti sinusoidali** (o di qualsiasi forma d'onda) **in ingresso vengono trasformati in una combinazione lineare di tensioni e correnti sinusoidali** (o di qualsiasi forma d'onda) **in uscita**, conservando la forma del segnale, mentre i secondi producono **in uscita delle tensioni e delle correnti che sono la sovrapposizione di diverse sinusoidi** (o di qualsiasi forma d'onda) perdendo il carattere originario del segnale in ingresso.

LA FISICA DEI SEMICONDUTTORI

I dispositivi elettronici basano il loro funzionamento sulle **proprietà di una particolare classe di materiali, i semiconduttori**. Il nome semiconduttore **non deve far pensare che essi si comportino quasi come metalli conduttori**, dal momento in cui **i due non condividono quasi nessuna proprietà**; un semiconduttore è tale perché, sotto opportuni processi produttivi, **permette l'aumento della propria conducibilità in un intervallo di circa 10 ordini di grandezza** e perché è **caratterizzato dalla presenza di cariche mobili positive, le lacune, in aggiunta a quelle negative, gli elettroni**, che sono l'unica carica mobile dei metalli.

Il **silicio** è uno dei semiconduttori più utilizzati per la produzione di dispositivi elettronici, grazie alle sue proprietà da semiconduttore e grazie al fatto che è **uno degli elementi più abbondanti che esistono in natura**, basti pensare che il 90% della crosta terrestre è fatta di silicio; tuttavia, **il silicio viene spesso trovato sotto forma di silicati**, mentre **in elettronica si lavora con la sua forma cristallizzata**, ottenuta a partire da un nucleo già cristallizzato attorno al quale è fatto ruotare a velocità basse il silicio ad alta temperatura che deve cristallizzare. La necessità della forma cristallina per i circuiti elettronici risiede nel fatto che essa **garantisce una struttura atomica perfettamente ordinata e periodica**:



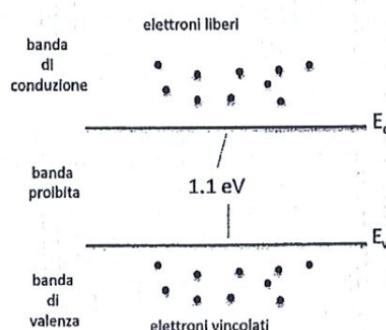
Il silicio è un atomo **tetravalente**, cioè con **quattro elettroni spaiati sul guscio di valenza**, il che permette alla struttura cristallina un **perfetto equilibrio tra le cariche**; infatti, **ogni atomo della struttura è legato con altri quattro atomi di silicio** con i quali condivide due elettroni ognuno in un legame covalente, che garantisce notevole stabilità. Di conseguenza, **il silicio cristallizzato appare un materiale elettricamente neutro** e privo di alcuna possibilità di conduzione; in realtà, per quanto forte possa essere, **il legame covalente può essere spezzato fornendo ad un elettrone una quantità sufficientemente grande di energia**.

Uno dei fattori che può contribuire alla liberazione di un elettrone da un legame covalente è la **distribuzione dell'energia tra i vari elettroni**; ad una certa temperatura, **un qualsiasi materiale è soggetto ad un movimento vibratorio microscopico** (non percepibile ai nostri occhi) che dipende **dalla temperatura** stessa e, essendoci **in un cm^3 di silicio circa 10^{22} atomi di silicio** e un numero

di elettroni pari a quattro volte questo valore, si può ben tenere conto che **l'energia media di tutti gli elettroni non è l'energia posseduta da ogni singolo elettrone**. Segue che in un blocco di silicio ci possono essere alcuni elettroni che hanno un energia sufficientemente inferiore rispetto alla media che devono bilanciare alcuni elettroni che hanno una duale quantità di energia superiore alla media che gli permette di divincolarsi dai legami covalenti. Questi elettroni diventano elettroni liberi e la loro concentrazione, in un cm^3 di silicio, è circa di 10^{10} , detta **concentrazione intrinseca** (n_i). La relazione tra la concentrazione intrinseca e la temperatura è mediata dalla costante di Boltzmann ($k = 8.6 \cdot 10^{-5} \text{ eV/K}$):

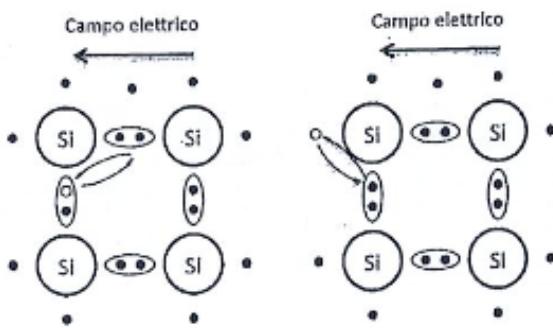
$$n_i \approx 1.25 \cdot 10^{20} \left(\frac{T}{300} \right)^{\frac{3}{2}} e^{-\frac{0.6}{kT}}$$

In un qualsiasi atomo, un elettrone o è strettamente legato al suo nucleo o è libero di muoversi; nel primo caso l'energia dell'elettrone energia apparterrà ad un intervallo di valori relativamente bassi, detta **banda di valenza** (il cui estremo superiore è E_V), mentre se è libero significa che è stata fornita un'energia superiore ad una certa soglia E_C , che è l'estremo inferiore della banda di conduzione. Nei semiconduttori accade che $E_C > E_V$, individuando una **banda di energie** (detta **banda proibita**) a cui nessun elettrone può appartenere; segue che per liberare un elettrone da un legame covalente serve fornirgli un'energia pari almeno a E_V , che equivale a dire sottrarre da uno o più elettroni la stessa quantità di energia. Si ricorda che nel mondo quantistico l'energia è una grandezza non continua ma distribuita in pacchetti, detti quanti di energia.



La dimensione della banda proibita è inversamente proporzionale alla conducibilità elettrica del materiale ed è caratteristica per ogni classe: **nei conduttori è molto bassa**, per permettere un facile spostamento, **negli isolanti è estremamente elevata**, per impedire lo spostamento, e **nei semiconduttori è né troppo bassa né troppo elevata**, l'importante è che sia **modulabile mediante appositi processi produttivi**.

Nel silicio, l'energia da fornire per spostare un elettrone dalla banda di valenza alla banda di conduzione è di circa 1.1 eV; a patto di fornire una tale energia, l'elettrone condiviso nel legame covalente viene divincolato e lascia dietro di sé una lacuna, cioè un legame covalente non soddisfatto che rappresenta una carica elettrica positiva. Poiché si tratta di posizioni energeticamente equivalenti, è probabile che in seguito all'agitazione termica la lacuna venga occupata in maniera casuale dagli elettroni nei legami covalenti adiacenti, spostando ulteriormente la lacuna in una direzione altrettanto casuale. Se ai capi del blocco di silicio cristallizzato in questione fosse posta una differenza di potenziale in grado di generare campo elettrico, gli elettroni che andrebbero a riempire un'eventuale lacuna non proverebbero più da direzioni casuali ma dalla direzione del campo elettrico, spostando le lacune nel verso puntato da quest'ultimo in un movimento fittizio di cariche positive.



In relazione alla figura appena mostrata, dove il campo elettrico è diretto verso sinistra, è ragionevolmente più probabile che la posizione libera della lacuna venga occupata da un elettrone proveniente da sinistra; in queste condizioni, il campo elettrico sovrappone all'agitazione termica un moto preferenziale.

Dal punto di vista quantistico (sebbene non si approfondisca più di tanto questa posizione) **la lacuna assume il ruolo di carica elettrica positiva libera** e, sebbene sia stata definita in relazione al movimento degli elettroni, considerarla come **il movimento opposto di questi ultimi risulta in gravi errori macroscopici; le due cariche sono indipendenti**, sebbene la loro definizione sia legata. Nel momento in cui un elettrone si libera da un legame covalente si genera una lacuna ed è ragionevole pensare che **la concentrazione di lacune sia direttamente proporzionale a quella di elettroni**:

$$n = n_i = p$$

Dove **n e p sono le concentrazioni di elettroni e lacune**, che si egualgiano in numero alla concentrazione intrinseca del silicio. Un semiconduttore di questo tipo prende il nome di **semiconduttore intrinseco (o puro)** e vi si può dimostrare facilmente che:

$$n \cdot p = n_i^2$$

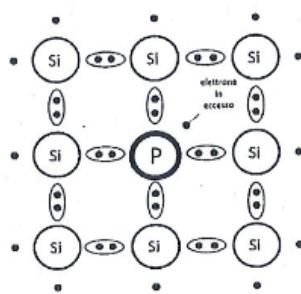
Questo risultato prende il nome di **legge dell'azione di massa** e porta con sé importanti considerazioni in relazione ai semiconduttori estrinseci.

La **conducibilità elettrica** di questo materiale è:

$$\sigma = qKn_i \approx 10^{-6} \frac{\text{S}}{\text{cm}}$$

Che è **estremamente bassa**, tanto che si può considerare il silicio intrinseco come un isolante; tuttavia, la comodità dei semiconduttori risiede nella possibilità di **modulare a piacere la conducibilità elettrica inserendo delle impurità di natura opportuna**, detti **droganti**, che alterano le proprietà elettriche ma non chimiche del silicio. Questo tipo di silicio è detto **semiconduttore estrinseco (o drogato)**.

Si prenda in considerazione un semiconduttore di silicio **drogato con atomi di fosforo (P)**, appartenenti al **V gruppo** e dotati di **cinque elettroni sul guscio di valenza**, contro i quattro del silicio. Questa discrepanza sul guscio di valenza fa sì che **il quinto elettrone del fosforo non sia vincolato in alcun legame covalente**, rendendolo una **carica libera** (l'energia che lo lega al fosforo è di soli 0.44 eV); si può dire, quindi, che **l'atomo di fosforo ha donato al reticolo cristallino di silicio un elettrone**, rendendolo un **atomo donatore**.

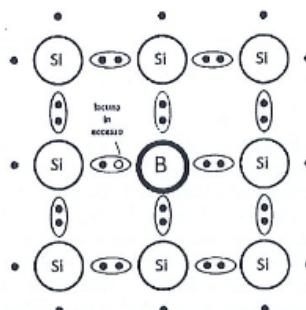


In genere, **tutti gli atomi del V gruppo possono essere atomi donatori**, sebbene ci sia industrialmente un limite di dimensione (finalizzato alla non alterazione della struttura geometrica del reticolo cristallino) che fa preferire gli atomi di fosforo (P), arsenico (As) e antimonio (Sb) al drogaggio donatore. Appare chiaro che **ad ogni atomo donatore "sciolto" nel silicio corrisponde un elettrone libero di conduzione** che si aggiunge alla normale popolazione generata termicamente; se, per esempio, viene sciolta una quantità pari a 10^{15} atomi/cm³ si ottiene un proporzionale aumento di elettroni liberi nel reticolo cristallino, aumentando di cinque ordini di grandezza la loro concentrazione. Ci si chiede se l'inserimento di queste impurità possa alterare la natura chimica e strutturale del semiconduttore; in realtà, **tenendo in considerazione che la concentrazione del silicio è di circa 10^{22} atomi/cm³, inserire 10^{15} atomi/cm³ non va ad alterare il materiale dal punto di vista chimico** (si troverà **un atomo di drogante ogni 10,000,000 di atomi di silicio**). Le tecnologie di microelettronica ad oggi disponibili permettono di **modulare la quantità del drogaggio tra 10^{12} e 10^{20} atomi/cm³**, modificando la conducibilità elettrica a piacimento.

Un drogaggio di questo tipo prende il nome di **drogaggio di tipo N e la concentrazione di atomi donatori** (e quindi di elettroni liberi) **introdotta viene indicata con N_D** . È importante notare che, **in assenza di campi elettrici, il materiale drogato è ancora complessivamente neutro**: ad ogni elettrone divincolato corrisponde il relativo **atomo drogante che si è ionizzato positivamente**.

Un'altra considerazione importante coinvolge il **processo di generazione termica** (la rottura dei legami covalenti): nel semiconduttore drogato **avviene ugualmente** questo processo **ma in quantità e misure minori**, dal momento in cui **nella banda di conduzione sono già presenti degli elettroni che rendono più difficile trovare spazio per gli elettroni nella banda di valenza**. Inoltre, visto che **ad un elettrone fornito dal drogante non corrisponde alcuna lacuna**, in un semiconduttore estrinseco **la concentrazione di lacune non è la stessa di quella di elettroni**, come avveniva nei semiconduttori intrinseci; in particolare, **in un drogaggio di tipo N ci saranno molti più elettroni che lacune**, pur lasciando valida la legge dell'azione di massa.

Un **discorso duale** può essere fatto con la possibilità di **drogare il silicio per aumentare la concentrazione di lacune**; in questo caso andrà scelto come **agente drogante un atomo del III gruppo**, avente **tre elettroni nel guscio di valenza** (contro i quattro del silicio) per i quali **un legame covalente sarà privo di un elettrone**.



Per un elettrone proveniente da un legame vicino, **andare ad occupare la posizione lasciata libera è energeticamente poco dispendioso** (ci vogliono 0.045 eV), innescando potenzialmente un **movimento di lacune**. In questa configurazione, **il boro accetta un elettrone**, e pertanto è detto **atomo accettore**; un semiconduttore drogato con un atomo del III gruppo è detto **semiconduttore di tipo P** e la concentrazione di atomi accettori è indicata con N_A . Analogamente al precedente tipo di droggaggio, un semiconduttore estrinseco di questo tipo è **ancora un materiale elettricamente neutro in assenza di campi elettrici: per ogni lacuna che si forma c'è un atomo di boro ionizzato negativamente per l'accezione dell'elettrone**.

Si possono sviluppare in maniera duale per i semiconduttori di tipo P le stesse considerazioni già presentate con i semiconduttori di tipo N, come la modulazione della concentrazione di atomi accettori o l'aumento di lacune per lasciare invariata la legge dell'azione di massa. A causa di questa discrepanza tra concentrazione di lacune e di elettroni, **si possono individuare due tipologie di portatori nei due tipi di droggaggio**, portatori **maggioritari** e portatori **minoritari**:

	Portatori maggioritari	Portatori minoritari
Semiconduttori di tipo N	Elettroni	Lacune
Semiconduttori di tipo P	Lacune	Elettroni

Può capitare (come nei diodi) che **un semiconduttore sia drogato sia con atomi donatori che con atomi accettori**. In questa particolare configurazione nel materiale si trovano quattro tipi di cariche elettriche: gli **elettroni liberi introdotti dagli atomi donatori**, le lacune introdotte dagli atomi accettori, gli **ioni positivi costituiti dagli atomi donatori** che hanno perso un elettrone e gli **ioni negativi formati dagli atomi accettori** che hanno catturato un elettrone; tuttavia, **non tutte queste cariche sono cariche di movimento** (di conduzione), lo sono **solamente le lacune e gli elettroni** visto che **gli ioni sono saldamente legati al reticolo cristallino**. Poiché nel complesso il semiconduttore deve rimanere neutro, **l'insieme delle cariche introdotte deve comunque bilanciarsi perfettamente**:

$$p + N_A = n + N_D$$

Questa **legge di neutralità**, in seno al fatto che gli ioni sono cariche fisse ed elettroni e lacune di conduzione, **non implica la neutralità locale del semiconduttore**: si possono **creare degli squilibri locali di carica con una conseguente differenza di potenziale**; tuttavia, visto che non esiste energia gratuita, **integralmente la differenza di potenziale è nulla ed inutilizzabile ingegneristicamente**.

LE CORRENTI ELETTRICHE NEI SEMICONDUTTORI

Il titolo del capitolo inganna e fa intuire la presenza di **diverse correnti elettriche che si possono osservare in un semiconduttore estrinseco**, ognuna delle quali origina da un fenomeno diverso.

Poiché **nei semiconduttori sono presenti cariche libere**, è possibile **innescare meccanismi di conduzione che portano al loro movimento**; in generale, **ad una data popolazione di n particelle di carica q che si muovono lungo una determinata direzione con velocità v , è associato un vettore densità di corrente di modulo:**

$$J = \frac{I}{S} = qn\nu_q$$

In questa formulazione **non compare alcun'informazione sul portatore e sul campo elettrico da cui la velocità dei portatori è stimolata** perché sono **insiti nella definizione di velocità stessa**:

$$\nu_q = \mu_q \mathcal{E} = \left[\frac{cm^2}{s \cdot V} \right] \left[\frac{V}{cm} \right]$$

Dove μ_q è la **mobilità**, un parametro che **descrive l'inclinazione delle particelle a muoversi quando sollecitate da un campo elettrico \mathcal{E}** . Distinguendo lacune ed elettroni (che sono le cariche mobili), in un semiconduttore a temperatura ambiente:

$$1500 \frac{cm^2}{s \cdot V} = \mu_n > \mu_q = 500 \frac{cm^2}{s \cdot V}$$

$$\mu_n \approx 2,5(\sqrt{3})\mu_p$$

Gli elettroni sono più inclini a muoversi delle lacune perché non sono vincolati, come queste ultime, ai nuclei degli atomi. Segue che, in un semiconduttore sottoposto a campo elettrico, si innescano due moti di cariche, indipendenti e sovrapposti; a quello degli elettroni è associato un vettore densità di corrente, detta **corrente di drift**, di modulo:

$$J_{n-drift} = qn\mu_n \mathcal{E}$$

E a quello delle lacune:

$$J_{p-drift} = qp\mu_p \mathcal{E}$$

Entrambe le correnti hanno lo stesso verso del campo elettrico, dal momento in cui **il segno negativo della carica degli elettroni annulla il verso negativo della velocità** dovuto al movimento in direzione opposta al campo elettrico; segue che **la corrente di drift totale è la somma delle due componenti**:

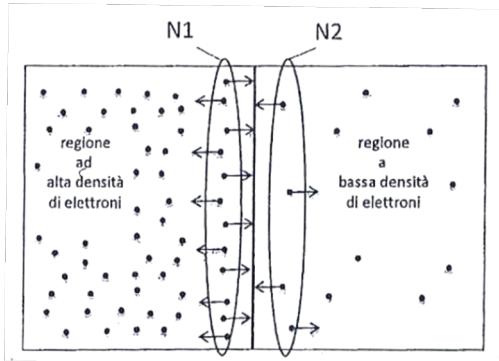
$$J_{drift} = J_{n-drift} + J_{p-drift} = qn\mu_n \mathcal{E} + qp\mu_p \mathcal{E}$$

Questa corrente è detta anche **corrente ohmica**, in quanto **rappresenta l'espressione locale della legge di Ohm**:

$$J_{drift} = (qn\mu_n + qp\mu_p) \mathcal{E} = (\sigma_n + \sigma_p) \mathcal{E} = \sigma \mathcal{E}$$

Infatti, **la conducibilità del silicio è data dalla somma delle conducibilità degli elettroni e delle lacune**.

Per comprendere l'esistenza del secondo tipo di corrente nei semiconduttori, **si supponga che** (senza chiederci il perché per il momento) **in una determinata regione di spazio ci sia una concentrazione N_1 di elettroni molto più alta che in un'altra a concentrazione $N_2 < N_1$** :



Si supponga, inoltre, che non sia applicata alcuna differenza di potenziale ai capi di questo materiale e che il movimento delle cariche sia dovuto unicamente all'agitazione termica. Poiché il moto di queste particelle è browniniano, si può ragionevolmente supporre che la probabilità delle cariche di andare verso destra o verso sinistra sia la stessa; tuttavia, è maggiore la probabilità che a valicare il confine tra le due regioni sia una carica appartenente alla regione N_1 . Supponendo comunque quest'ultimo evento equiprobabile al suo duale, si avranno $N_1/2$ portatori che si spostano da sinistra verso destra e $N_2/2$ che si spostano da destra verso sinistra: nella sezione di semiconduttore considerata si avrà uno spostamento maggiore verso destra e verrà generata una corrente di elettroni diversa da zero. Questa corrente prende il nome di **corrente di diffusione** e si osserva ogni volta c'è una disuniformità nella concentrazione delle cariche elettriche (i portatori liberi hanno una tendenza naturale ad andare dalle regioni a maggior concentrazione alle regioni a minor concentrazione, in un movimento entropico).

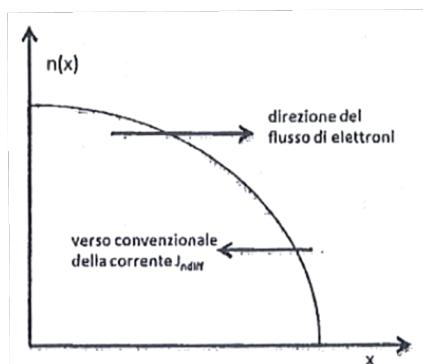
Formalmente, la corrente di diffusione dipende dalla concentrazione degli elettroni in funzione della posizione; in particolare dal suo gradiente (supposto monodimensionale per semplicità):

$$J_{n-diff} = qD_n \frac{dn}{dx}$$

Dove D_n è la **diffusività** (o coefficiente di diffusione degli elettroni) ed è un parametro **dipendente dalla temperatura del semiconduttore** (25mV a temperatura ambiente, 300K):

$$D_n = \mu_n V_t = \mu_n \frac{kT}{q} = \left[\frac{cm^2}{s} \right]$$

Con k costante di Boltzmann, T temperatura e $V_t = kT/q$ tensione termica; questa formula prende il nome di **relazione di Einstein**. Il segno positivo della corrente di diffusione è dovuto al fatto che il segno negativo delle cariche è bilanciato dalla convenzione per cui si è assunto come positivo il verso nel quale la concentrazione diminuisce, conducendo ad una derivata negativa.



Un discorso del tutto analogo può essere ripetuto per le lacune, dove:

$$J_{p-diff} = -qD_p \frac{dp}{dx}$$

Il segno negativo deriva dal fatto che la derivata continua ad essere negativa ma le cariche sono positive.

In definitiva, se nel semiconduttore è presente un gradiente nella concentrazione dei portatori di carica, allora è presente una corrente di diffusione data dalla somma delle correnti di diffusione degli elettroni e delle lacune:

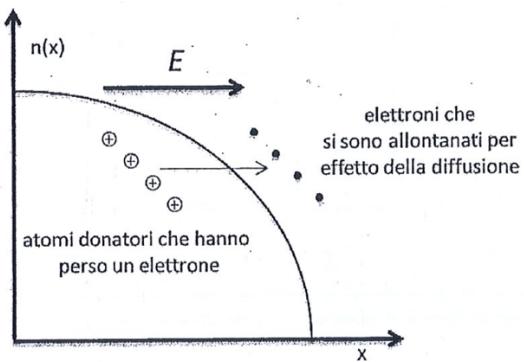
$$J_{diff} = qD_n \frac{dn}{dx} - qD_p \frac{dp}{dx}$$

Si può notare che, a differenza della corrente di drift, la corrente di diffusione non è dipendente da alcun campo elettrico. Nel caso in cui il semiconduttore sia soggetto ad una differenza di potenziale, si presentano sia correnti di drift che di diffusione:

$$J = J_{n-drift} + J_{n-diff} + J_{p-drift} + J_{p-diff} = q\mu_n n\varepsilon + qD_n \frac{dn}{dx} + q\mu_p p\varepsilon - qD_p \frac{dp}{dx}$$

Le condizioni di disuniformità che conducono ad una corrente di diffusione sono da ricercare in un drogaggio non uniforme, in cui la concentrazione N_D varia in funzione della posizione. Il silicio utilizzato nei dispositivi elettronici viene realizzato a partire da un particolare tipo di sabbia, la quarzite, sottoforma di ossido SiO_2 ; dopo la purificazione del composto, il silicio fuso, a cui è stato già aggiunto il drogante, viene fatto cristallizzare attorno ad un nucleo già cristallizzato mediante una rotazione particolarmente lenta da cui verrà prodotto un lingotto di semiconduttore di tipo P o N. Il drogaggio risultante è uniforme e per renderlo disuniforme si possono utilizzare due tecniche, una delle quali (la diffusione termica) prevede l'esposizione del silicio ad alte temperature in corrispondenza di droganti allo stato gassoso: intorno ai mille gradi le particelle di gas penetrano all'interno dei materiali e, a fronte di uno spessore del wafer di circa $300\mu m$, modificano il drogaggio nei primi micron o frazioni di esso, quindi sulla superficie; in alternativa, si potrebbe usare la tecnica dell'impantazione ionica, con la quale il drogante viene sparato all'interno del silicio da acceleratori di particelle, o il processo fotolitografico, per il quale una determinata geometria è trasferita sul silicio tramite una maschera ed un fotoresistore sottoposti a radiazione ultravioletta (questa tecnica è usata perlopiù in serie ad una delle due precedenti)

Poiché a valle è stato supposto il materiale in assenza di differenze di potenziale, è necessario che la corrente totale circolante sia nulla, non si può avere energia gratuita. Si ricordi che quando un elettrone (o una lacuna) si libera, l'atomo dal quale si allontana diventa uno ione di segno opposto che rimane vincolato al reticolo cristallino; ciò significa che, quando gli elettroni si spostano verso destra per effetto della diffusione, lasciano di sé un egual numero di ioni positivi, facendo manifestare nel materiale uno squilibrio di cariche con la conseguente nascita di un campo elettrico \mathcal{E} , il cui verso tende a riportare verso sinistra gli elettroni in questione.



In assenza di tensioni esterne applicate, la corrente di diffusione viene bilanciata da una corrente di trasporto di segno contrario prodotta dal campo elettrico appena rilevato; cioè, si stabilisce un equilibrio dinamico tra correnti di diffusione e di drift:

$$q\mu_n n \mathcal{E} + qD_n \frac{dn}{dx} = 0$$

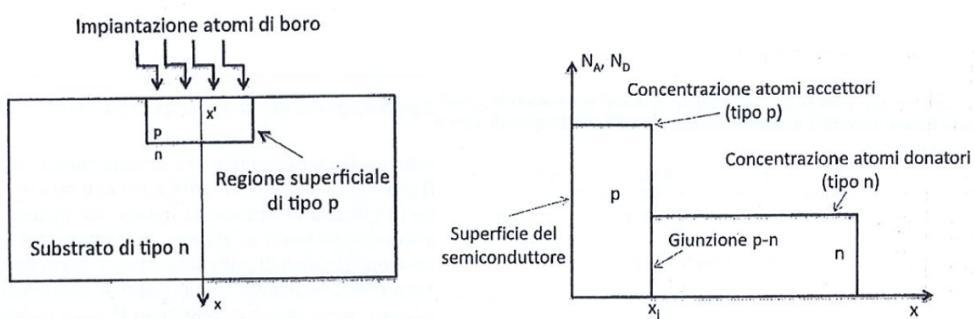
$$\mathcal{E} = -\frac{D_n}{\mu_n n} \frac{dn}{dx}$$

Dualmente per le lacune:

$$\mathcal{E} = \frac{D_p}{\mu_p p} \frac{dp}{dx}$$

LA GIUNZIONE PN

Dopo aver visto come è possibile modulare il droggaggio in uno stesso semiconduttore in funzione della posizione, si voglia considerare un dispositivo del genere:



Ottenuto dall'impiantazione in una fetta di silicio drogata di tipo N di atomi accettori (come il boro) in modo tale da rendere una parte della superficie del semiconduttore drogata di tipo P. Procedendo lungo l'asse x a partire dalla superficie e procedendo verso l'interno, si può incontrare dapprima una regione di semiconduttore di tipo P, caratterizzato da un'assegnata concentrazione (uniforme e costante) N_A di atomi accettori, per poi arrivare alla zona di tipo N caratterizzata da un'assegnata concentrazione (uniforme e costante) N_D di atomi donatori; la regione di transizione tra la zona drogata di tipo P e la zona drogata di tipo N viene chiamata giunzione metallurgica, o giunzione PN. La maggior parte dei dispositivi elettronici è costruita a partire da questo tipo di

struttura e la comprensione degli eventi che avvengono a ridosso della giunzione metallurgica è essenziale alla comprensione del funzionamento elettrico di suddetti dispositivi.

Le ipotesi sotto cui viene effettuato lo studio della giunzione PN sono quelle per cui **le concentrazioni di atomi donatori e accettori sono costanti** e per cui **il passaggio da una zona all'altra avviene in maniera netta**, ovvero nelle condizioni di **giunzione brusca**.

Poiché ad ogni atomo drogante corrisponde una carica libera, nella regione di tipo P il numero di lacune sarà pari al numero di atomi accettori e nella regione di tipo N il numero di elettroni sarà pari al numero di atomi donatori:

$$p(x < x_j) = N_A \wedge n(x > x_j) = N_D$$

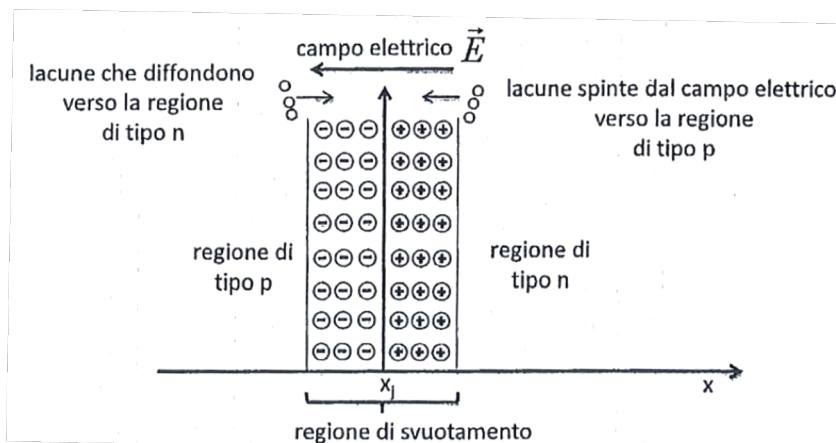
Per i portatori minoritari di ogni regione (elettroni nella regione P e lacune nella regione N) la concentrazione segue la legge dell'azione di massa:

$$n(x < x_j) = \frac{n_i^2}{N_A} \wedge p(x > x_j) = \frac{n_i^2}{N_D}$$

Ricordando che $n_i \approx 10^{20} \text{ cm}^{-3}$ e che i valori fisicamente realizzabili di N_A e N_D variano tra i 10^{13} e 10^{18} atomi/ cm^{-3} , si può concludere che:

$$p(x < x_j) \gg p(x > x_j) \wedge n(x > x_j) \gg n(x < x_j)$$

Ovvero che la concentrazione delle cariche libere è fortemente variabile lungo la profondità del semiconduttore e, pertanto, si instaura una naturale tendenza degli elettroni e delle lacune ad invadere le regioni dove sono presenti in più bassa concentrazione (processo di diffusione); in particolare, questa tendenza si traduce in un movimento di elettroni dalla regione N alla regione P e un movimento di lacune in verso opposto, risultando in due correnti di carica che si sommano nello stesso verso dell'asse x . Tuttavia, non è stata applicata alcuna tensione ai capi del semiconduttore, quindi la corrente totale deve essere nulla; cioè, ci deve essere lo stesso meccanismo di bilanciamento precedentemente mostrato: le lacune lasciano dietro di sé nello spostamento degli ioni negativi fissi, allo stesso modo gli elettroni, e a ridosso della giunzione metallurgica non si trovano più cariche mobili, lasciando il posto ad uno squilibrio di cariche elettriche fisse.



Il meccanismo di trasferimento appena mostrato **non avviene uniformemente su tutta la superficie delle due regioni ma solo su una regione limitata**, detta **regione di svuotamento**. Lo squilibrio di

cariche che si palesa nella regione di svuotamento produce un campo elettrico con un verso tale da opporsi al movimento delle lacune, delle quali viene favorito il movimento di ritorno al lato P (analogamente per gli elettroni ma con versi opposti).

Dunque, in assenza di tensioni applicate, si raggiunge l'equilibrio: per ogni lacuna o elettrone che si allontana dalla propria regione per diffusione (corrente di diffusione) c'è una lacuna o un elettrone che viaggia in senso contrario sotto l'influenza del campo elettrico generato dalla diffusione stessa (corrente di drift). In termini matematici:

$$\begin{cases} J_p = J_{p-diff} + J_{p-drift} = 0 \\ J_n = J_{n-diff} + J_{n-drift} = 0 \end{cases}$$

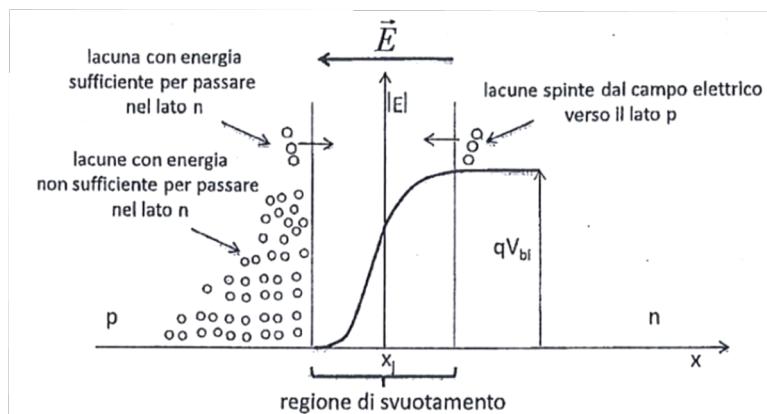
È interessante notare come le **correnti di drift di entrambe le regioni siano generate dai portatori minoritari**, quindi:

$$\begin{cases} J_{p-drift} = q\mu_p \frac{n_i^2}{N_D} E \\ J_{n-drift} = q\mu_n \frac{n_i^2}{N_A} E \end{cases}$$

Volendosi riferire solamente alle lacune per la possibilità di fare un discorso duale sugli elettroni, è possibile osservare il fenomeno anche da un punto di vista energetico. Il campo elettrico ostacola il moto diffusivo delle lacune e, dal punto di vista di queste ultime, può essere visto come una barriera che è possibile oltrepassare solo se si ha un'energia superiore a quella della barriera stessa. L'altezza (energetica) della barriera in funzione della posizione è valutata a partire dal campo elettrico:

$$V(x) = - \int \bar{E}(x) dx$$

Che corrisponde all'andamento del potenziale elettrico lungo il semiconduttore e, se vi è moltiplicata la carica q , all'andamento dell'energia potenziale associata ai portatori, che assume la forma:



Al di fuori della regione in cui è confinato lo squilibrio di cariche, cioè oltre la regione di svuotamento, è ragionevole pensare che il campo elettrico sia nullo (e quindi il potenziale costante). Dato il verso del campo elettrico si trova però, che l'energia potenziale associata alle lacune che si trovano nel lato N è maggiore in modulo all'energia potenziale delle lacune nel

lato P; la differenza di potenziale tra il lato N e il lato P è detta potenziale di built – in ed è legato alla concentrazione dei droganti che formano la giunzione:

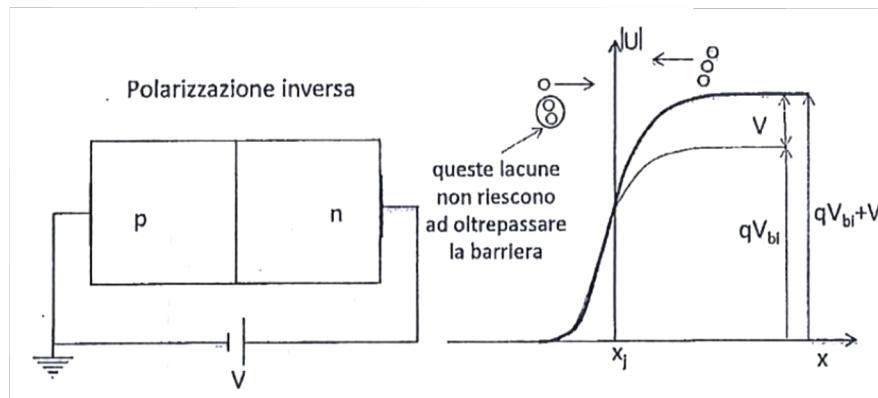
$$V_{bi} = V_t \ln \left(\frac{N_A N_D}{n_i^2} \right)$$

Le lacune che si trovano nel lato N possiedono un'energia potenziale qV_{bi} volte maggiore di quelle che si trovano nel lato P; questa quantità è detta **energia di barriera e rappresenta la quantità di energia da possedere per una lacuna per poter effettivamente migrare dal lato P al lato N (ovviamente va aggiunta all'energia elettrica quella cinetica).**

Come l'immagine può semplificare, **nel lato P statisticamente ci sarà una maggioranza di lacune con un'insufficiente quantità di energia per la migrazione; nel lato N, però, le lacune non incontrano alcun ostacolo, anzi sono favorite a discendere la barriera di potenziale e a spostarsi verso la regione P. In questa regione le lacune saranno dei portatori minoritari, quindi presenti in minor concentrazione, e sarà ragionevole pensare che le poche lacune che riescono a spostarsi da P a N sono equivalenti in numero alle stesse poche lacune che si spostano da N a P, rendendo la corrente netta che attraversa la giunzione nulla.** In riferimento agli elettroni è possibile sviluppare un **ragionamento duale**, arrivando ad ottenere le equazioni delle correnti precedentemente scritte.

Ci si chiede cosa succede se al dispositivo appena descritto viene applicata una tensione esterna che altera la condizione di equilibrio termodinamico e tra le correnti. Questo processo viene detto **polarizzazione della giunzione PN**.

Si supponga di applicare un generatore di tensione V con il terminale positivo applicato alla regione di tipo N e quello negativo alla regione di tipo P:

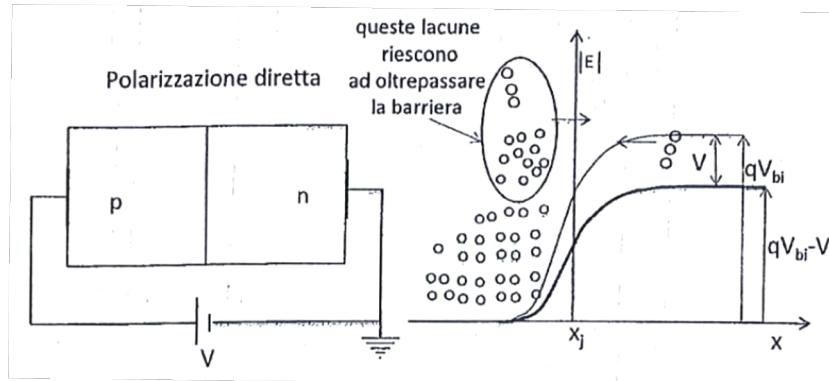


Questa condizione prende il nome di **polarizzazione inversa**. La tensione applicata dall'esterno è **concorde al potenziale di built – in**, andando a **sommarsi ad esso e ad incrementare l'altezza della barriera di energia**; questa circostanza **modifica l'equilibrio delle correnti** descritto precedentemente, andando a **ridurre il numero di lacune** presenti nel lato P che possiede energia sufficiente per superare la barriera, mentre le lacune che si trovano nel lato N continuano a non avere ostacoli e a poter facilmente migrare verso la regione P. Il numero di tali lacune non è cambiato, quindi **non cambia neanche la corrente (di drift) ad esse associata**; in altri termini, **in condizioni di polarizzazione inversa la corrente di diffusione** (legata all'altezza della barriera di energia) **diminuisce mentre la corrente di drift** (non legata all'altezza della barriera di energia) **non cambia**:

$$J_{p-diff} < J_{p-drift} \Rightarrow J_{tot} = J_{p-diff} - J_{p-drift} \neq 0$$

Questa corrente è negativa ed estremamente piccola (detta **ingegneristicamente nulla**, non usabile) perché legata alla concentrazione di portatori minoritari nella regione N; inoltre, visto che è prodotta principalmente dalla corrente di drift, non sarà dipendente dalla tensione applicata.

Si supponga di applicare un generatore di tensione V con il terminale positivo applicato alla regione di tipo P e quello negativo alla regione di tipo N:

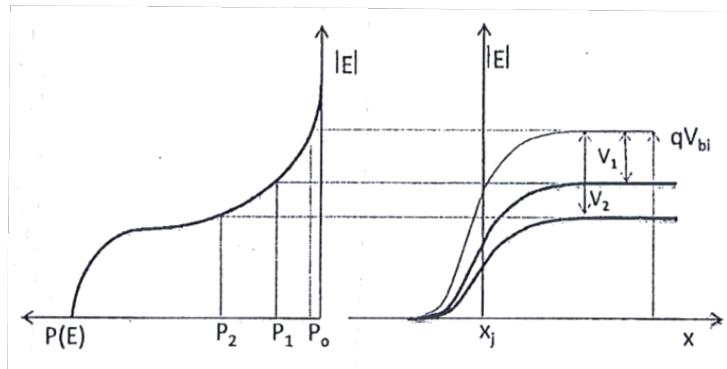


Il dispositivo si dice **polarizzato direttamente**. In questo caso il potenziale applicato è discordante alla tensione di built-in, sottraendosi ad esso ed abbassando la barriera di energia; in seno alle stesse considerazioni fatte nel caso della polarizzazione inversa, si riconosce che la corrente di drift non varia in funzione della barriera di energia ma che la corrente di diffusione aumenta, dal momento in cui è minore l'energia sufficiente a superare la barriera. Questa corrente non è indipendente dalla tensione applicata, più essa è alta e più si abbassa la barriera, trovando così più lacune con energia sufficiente.

$$J_{p-diff} > J_{p-drift} \Rightarrow J_{tot} = J_{p-diff} + J_{p-drift} > 0$$

Per ottenere un **legame quantitativo tra la tensione applicata e la corrente circolante** ci si può riferire (nella fisica dello stato solido) alla **probabilità di trovare una lacuna con energia inferiore ad E** :

$$P(E) = \frac{1}{1 + e^{\frac{E-E_F}{kT}}}$$



Dove E_F è un **parametro caratteristico dei semiconduttori**, noto come **livello di Fermi** ($P(E_F) = 1/2$). Si può notare come all'aumentare della tensione applicata diminuisce la barriera di energia e aumenta la probabilità che le lacune abbiano energia sufficiente ad oltrepassarla.

Poiché il numero di lacune coinvolte nella diffusione dipende in maniera esponenziale dalla tensione applicata e poiché il legame tra queste e la corrente è di diretta proporzionalità, allora anche la corrente circolante nel semiconduttore aumenterà in maniera esponenziale in funzione della tensione applicata:

$$J_{p-diff} \propto J_o e^{\frac{V}{V_t}}$$

Dove si può dimostrare che J_o è la stessa corrente che circola in condizioni di polarizzazione inversa, detta **corrente di saturazione inversa della giunzione**. Con la seguente relazione è possibile descrivere (quasi perfettamente) la legge di variazione della corrente in funzione di qualsiasi tensione, positiva o negativa, applicata ai capi di una giunzione PN:

$$I = I_0 \left(e^{\frac{V}{V_t}} - 1 \right)$$

Considerando $I_0 = |J_0|A$, con A sezione del dispositivo. Il modello predice con accuratezza la corrente nelle tre condizioni di funzionamento elencate in questo capitolo:

- **Condizione di equilibrio ($V = 0$)**

$$I = I_0(e^0 - 1) = 0$$

Non circola corrente, come rilevato attraverso le considerazioni sul campo elettrico.

- **Condizione di polarizzazione inversa ($V < 0$)**

$$I = I_0 \left(e^{\frac{V}{V_t}} - 1 \right) \approx -I_0 < 0$$

La corrente che circola è negativa e prossima allo zero, come rilevato attraverso le considerazioni sulla corrente di drift in polarizzazione inversa.

- **Condizione di polarizzazione diretta ($V > 0$)**

$$I = I_0 \left(e^{\frac{V}{V_t}} - 1 \right) > 0$$

La corrente che circola è positiva ed esponenziale, come rilevato attraverso le considerazioni sulla corrente di diffusione in polarizzazione diretta.

IL DIODO

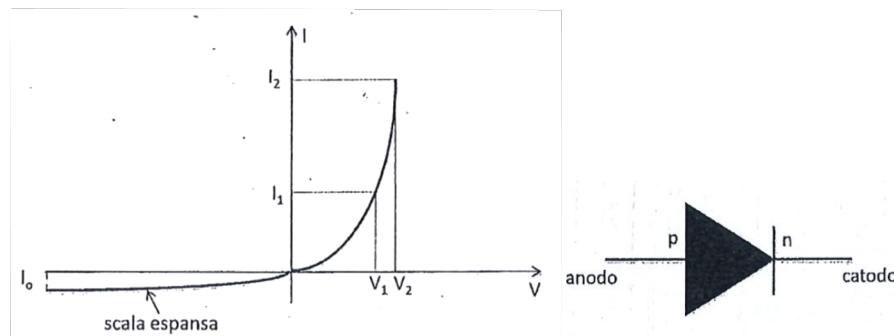
Il dispositivo elettrico composto da una giunzione PN è detto **diodo** e la sua **caratteristica** può essere modellata dalla legge della corrente in funzione della tensione che è stata precedentemente rilevata:

$$I = I_0 \left(e^{\frac{V}{V_t}} - 1 \right)$$

Ovviamente, il diodo è un dispositivo non lineare che presenta un comportamento molto diverso a seconda del fatto che la tensione applicata ai suoi capi sia positiva o negativa:

- **Tensione negativa**, la giunzione PN che compone il diodo è in **polarizzazione inversa** e nel dispositivo scorre una corrente ingegneristicamente nulla (il diodo in queste condizioni è interdetto);
- **Tensione positiva**, la giunzione PN che compone il diodo è in **polarizzazione diretta** e nel dispositivo per piccole variazioni di tensione corrispondono elevate variazioni di correnti;
- **Tensione nulla**, la giunzione PN è in **equilibrio** e nel dispositivo non circola corrente, rendendolo un bipolo passivo.

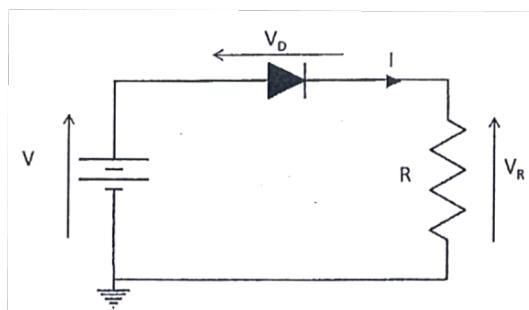
Si può, quindi, intuire che **il diodo funge da valvola per il passaggio della corrente**.



Quello in figura è il **simbolo circuitale del diodo**, rappresentato da un **catodo** in corrispondenza della **regione di tipo N** del semiconduttore e da un **anodo** in corrispondenza della **regione di tipo P**; la corrente può circolare (praticamente) solo dall'**anodo** al **catodo**, ovvero solo nel verso indicato dalla freccia.

Come si può osservare nel grafico corrente – tensione (IV), le tensioni per cui la corrente inizia ad impennarsi notevolmente sono nell'intervallo **0.6 – 0.7V**, detta **tensione di soglia**. Nella realtà, le funzioni esponenziali non hanno alcuna soglia e la definizione appena data dipende da una **pura sensazione visiva**, che può cambiare al variare della scala con cui si rappresenta la funzione; infatti, nel **data sheet** (targhe) di un diodo non è mai data la **tensione di soglia**, che va sempre **definita in relazione alla corrente tipica di utilizzo I_F** . La definizione più rigorosa di tensione di soglia prende in esame quel **valore per il quale la corrente del diodo supera un centesimo della corrente di utilizzo**, oppure l'intersezione con l'asse x della tangente alla curva nel punto corrispondente alla **corrente di utilizzo**.

Si consideri il seguente circuito:



Il **bilancio delle tensioni**, in seno alla LKT, è:

$$V = V_D + V_R$$

In seno alla LKC, si può affermare che **la corrente che circola nel circuito è unica**, unendo questi risultati alle caratteristiche dei bipoli:

$$V = V_t \ln\left(\frac{I}{I_0} + 1\right) + RI$$

Questa è **un'equazione trascendente nell'incognita I che non ammette soluzione analitica**. Il fatto che **un circuito di questa semplicità non sia risolvibile in maniera immediata** richiede l'esigenza di **trovare la possibilità di soluzioni approssimative**. Considerando $V = 12V$, $I_0 = 10^{-12}A$ e $R = 10\Omega$, si può ottenere come risultato della corrente $I = 1.1306A$; si noti che la caduta di tensione ai capi del resistore è:

$$V_R = 11.3V$$

Mentre quella ai capi del diodo:

$$V_D = 0.7V$$

Quindi, **il 94% della tensione di alimentazione cade sul resistore e solo il 6% sul diodo**. Il risultato è coerente con l'osservazione fatta per cui **la caduta di tensione ai capi di un diodo non varia significativamente in un range molto ampio di correnti** (e viceversa); in particolare, in un range di correnti molto ampie la tensione del diodo varia tra i $0.6V$ e i $0.8V$. Segue che, **in un'analisi di prima approssimazione, si può ritenere di conoscere la caduta sul diodo**, ponendola ad esempio **sempre uguale a $0.7V$** e supponendo che la corrente che lo attraversa sia imposta dagli altri elementi del circuito.

La regola generale che descrive la possibilità di effettuare un'analisi approssimativa dei circuiti in cui sono contenuti diodi è **contenuta nella formula che descrive il funzionamento del circuito precedente**:

$$I = \frac{V - V_D}{R}$$

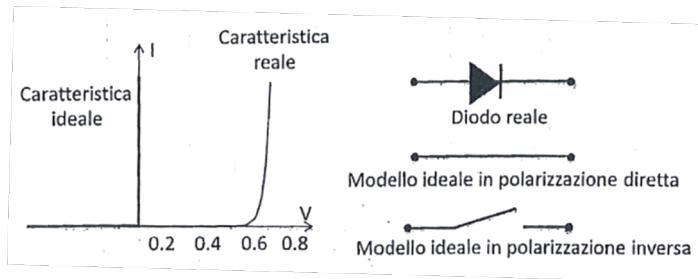
L'incognita è la tensione V_D ai capi del diodo. La possibilità di evitare la trattazione analitica dettagliata dipende dalla possibilità di ritenere nota la tensione V_D relativamente alla tensione V ; ovvero, **assumendo per certo che la tensione V_D possa assumere solo valori compresi tra $0V$ e $0.8V$, l'errore che si commette nel calcolo della corrente è ridotto solo se $V \gg 0.8V$** . Nel caso in cui si riconosca tale relazione, si può procedere ad una rappresentazione semplificata del circuito mediante modelli semplificativi del diodo.

Quando $V \gg V_D$, la semplificazione più immediata consiste nel trascurare completamente V_D rispetto a V e ritenere che la corrente circolante nel circuito sia imposta solo dalla resistenza:

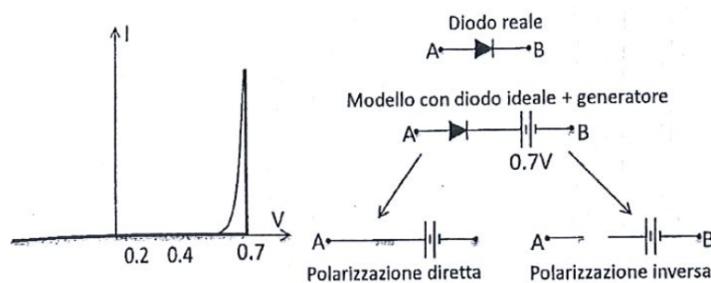
$$I = \frac{V}{R}$$

Utilizzando i valori dell'esempio precedente, si ottiene una corrente di $1.2A$, commettendo un **errore inferiore al 10%**. Un'approssimazione di questo tipo è **assolutamente accettabile nella dimensione di comprensione del circuito** (anche complesso), riservando l'analisi dettagliata solo all'effettivo dimensionamento dei componenti.

Il modello che descrive l'approssimazione appena effettuata prende il nome di **modello del diodo ideale** e permette di scrivere la **caratteristica del diodo come una caratteristica lineare a tratti**; infatti, ponendo $V_D = 0$ si considera il dispositivo come un cortocircuito per $V > 0$ e un circuito aperto per $V < 0$:



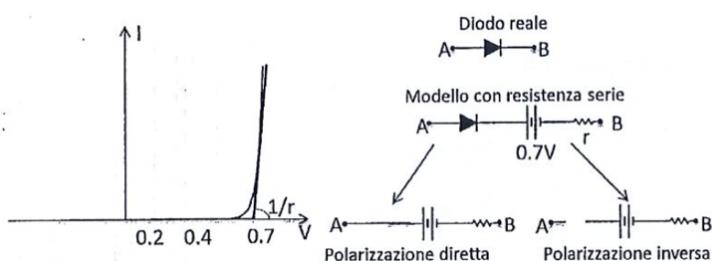
Rispetto alla semplificazione appena mostrata, **si ottengono risultati meno approssimativi** se si **imposta la tensione costante a 0.7V**; infatti, si è precedentemente notato che, in un ampiissimo range di correnti, la tensione sul diodo è compresa tra $0.6V$ e $0.8V$, rendendo ragionevole l'approssimazione al valor medio di questo intervallo. **Il modello che descrive questa approssimazione**, detto **modello a caduta di tensione costante**, prevede l'implementazione di **un generatore di tensione a 0.7V in serie ad un diodo ideale**:



In questo modo, **si avrà passaggio di corrente solo se la tensione fra i punti A e B supera i 0.7V**; in corrispondenza di questo valore viene sostituito il circuito aperto con il cortocircuito, attivando il generatore di tensione ed erogando la tensione in questione. In un circuito simile a quello usato precedentemente come esempio, la corrente sulla resistenza sarebbe data da:

$$I = \frac{V - 0.7}{R}$$

Questo modello non tiene in considerazione il fatto che, sebbene di poco, la tensione varia lo stesso al variare della corrente che scorre nel circuito. Per approssimare ancora al meglio la caratteristica del diodo si usa il **modello con resistenza serie**, che **linearizza la retta verticale introdotta dal precedente modello** attraverso un resistore in serie al generatore di tensione:

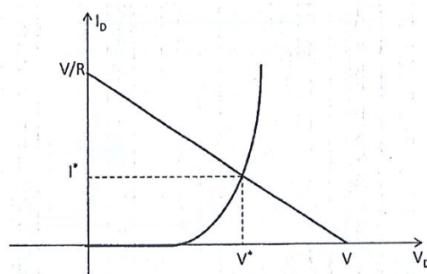


Il valore della resistenza più opportuno dipende dal livello di corrente che si immagina debba circolare nel diodo. Per questo motivo, il **modello con resistenza serie non è adatto ad una veloce analisi comportamentale del circuito;** infatti, aggiunge rispetto ai precedenti **delle complicazioni** che lo rendono **di scarso interesse pratico,** sebbene sia molto più vicino degli altri al **modello reale.**

Un modo alternativo per l'analisi dei circuiti usato spesso è il **metodo grafico**, che consiste semplicemente nella **risoluzione grafica del sistema di equazioni** che descrive il circuito. In relazione sempre allo stesso esempio:

$$\begin{cases} V = V_D + RI_D \\ I_D = I_0 e^{\frac{V_D}{V_t}} \end{cases}$$

In entrambe compaiono le incognite V_D e I_D del diodo. La rappresentazione della seconda equazione nel piano tensione – corrente (IV) è:

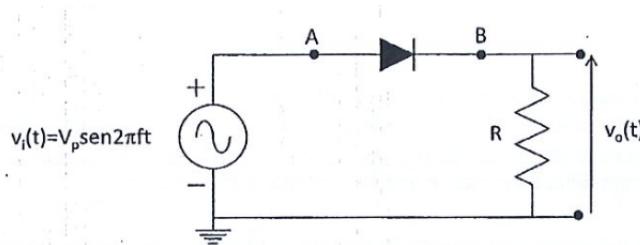


La prima equazione, invece, rappresenta una retta, detta **retta di carico**, che interseca l'asse x per $V_D = V$ e l'asse y per $I_D = V/R$; la **soluzione del sistema** di equazioni è rappresentata, graficamente, dall'intersezione delle due curve, permettendo di stabilire la corrente I^* e la tensione V^* che circolano nel circuito.

La potenza di questo metodo si palesa quando si è consapevoli che **un qualsiasi circuito collegato ad un diodo può essere sostituito con il circuito appena analizzato mediante l'applicazione del teorema di Thevenin.**

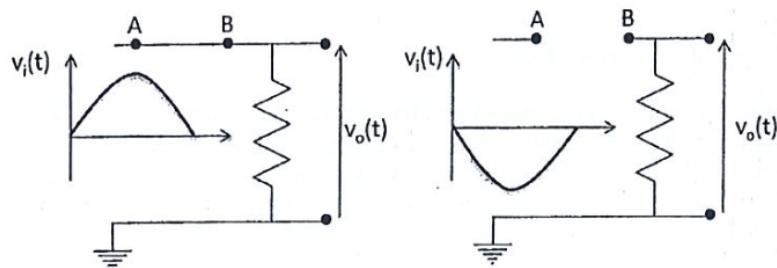
Essendo il dispositivo elettronico più semplice, **il diodo si presta bene in innumerevoli applicazioni**, in accoppiata con altri dispositivi elettronici o in varie combinazioni. Di seguito sono riportate **alcune applicazioni**, semplicissime ma molto diffuse, in cui il diodo svolge un ruolo fondamentale.

IL RADDRIZZATORE A SINGOLA SEMIONDA

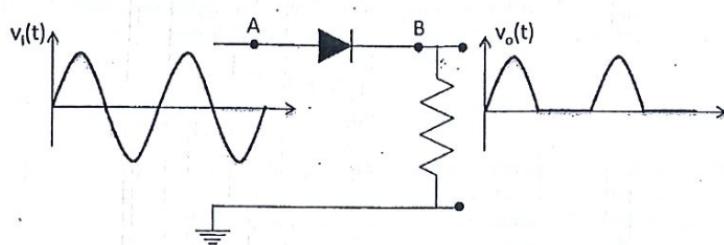


Questo circuito è, **topologicamente, identico a quello in precedenza mostrato come esempio**, con l'unica differenza dal punto di vista eziologico: **il generatore eroga una tensione $v_i(t)$ sinusoidale.** Si vuole ottenere la forma della tensione $v_o(t)$ ai capi del resistore.

Essendo la tensione di alimentazione una funzione del tempo, **il diodo non sarà necessariamente sempre in polarizzazione diretta o inversa**; è necessario studiare sotto quali condizioni si presenta una o l'altra polarizzazione. Osservando il circuito, si può notare che nel semiperiodo in cui la tensione presenta la semionda positiva il diodo si trova in polarizzazione diretta (e sarà possibile sostituirlo con un cortocircuito), mentre nel semiperiodo in cui la tensione è negativa il diodo si trova in polarizzazione inversa (e sarà possibile sostituirlo con un circuito aperto):

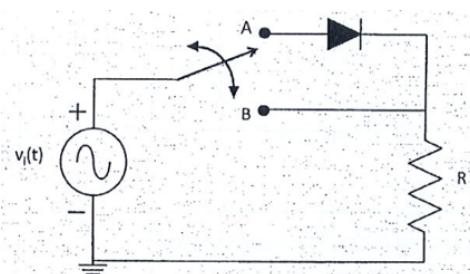


Quindi, nel semiperiodo positivo (a cui è associata la semionda positiva) ai capi della resistenza viene trasmessa integralmente la tensione di ingresso, mentre nel semiperiodo negativo (a cui è associata la semionda negativa) ai capi della resistenza non si rileva più alcuna tensione. Poiché la tensione in ingresso è sinusoidale, quindi periodica, anche questo meccanismo di alternanza tra tensione di uscita sinusoidale e nulla sarà periodico; in particolare, le due tensioni avranno lo stesso periodo.



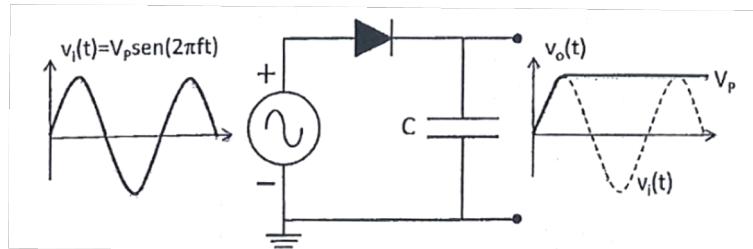
Il prodotto di questo circuito è una forma d'onda che elimina le semionde negative della tensione sinusoidale in ingresso pur conservando quelle positive; per questo motivo è definito raddrizzatore a singola semionda. I raddrizzatori a singola semionda erano largamente impiegati negli asciugacapelli, attraverso un interruttore che commuta un diodo o un cortocircuito ad un resistore riscaldante; il risultato è la modulazione della potenza dissipata sulla resistenza per un fattore di $\frac{1}{2}$ a causa del dimezzamento dei punti di funzionamento a tensione non nulla:

$$P_A = \frac{V_{eff}^2}{R} \wedge P_B = \frac{V_{eff}^2}{2R}$$

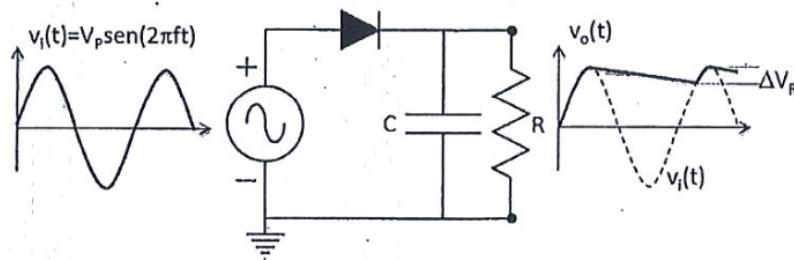


IL RILEVATORE DI PICCO

Nel circuito in figura, supponendo il condensatore inizialmente scarico, la tensione ai capi del condensatore segue la tensione di alimentazione nel primo quarto di periodo, in cui il diodo è polarizzato direttamente; una volta arrivato all'apice della semionda, il condensatore non può più scaricarsi, dal momento in cui farlo significherebbe riversare una corrente di scarica verso il generatore, attraversando il diodo in polarizzazione inversa.



Quindi, una volta che la tensione ha raggiunto il suo valore di picco e inizia a scendere, sul condensatore la tensione è ancora pari al valore di picco ed il diodo si trova contropolarizzato, isolando il condensatore. Il circuito si occupa di trasformare una tensione sinusoidale in una tensione continua ma solo se il condensatore resta isolato; infatti, collegando in uscita un utilizzatore al quale si vuole fornire la tensione continua in questione, il condensatore, durante la fase in cui il diodo è interdetto, si scarica sulla resistenza fintantoché la tensione di ingresso non supera quella ai capi del condensatore:

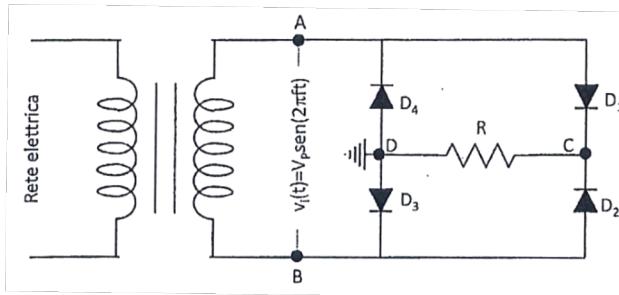


La scarica del condensatore avviene con costante di tempo RC e la forma d'onda presenta delle periodiche discese lineari seguite da previ risalite sinusoidali; queste oscillazioni prendono il nome di **ripple** e sono tanto minori quanto più grande è il prodotto RC .

IL RADDRIZZATORE A DOPPIA SEMIONDA

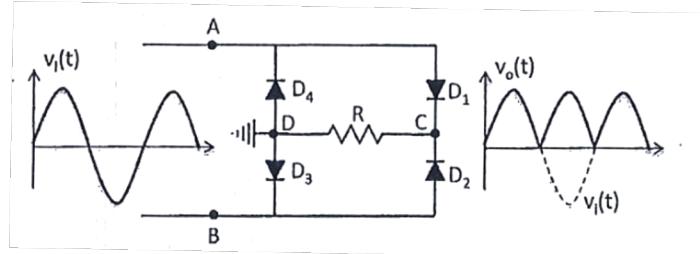
Nel circuito seguente, detto raddrizzatore a doppia semionda, si può notare che il generatore di tensione sinusoidale è stato sostituito da un trasformatore; il motivo risiede nel fatto che i dispositivi che usano questi circuiti ricevono tensioni relativamente piccole rispetto a quelle di rete (di valore efficace 220V) e necessitano di trasformatori per abbassarne il valore. Il sistema composto da trasformatore e raddrizzatore (quindi che abbassa e rende continua la tensione) è detto **alimentatore**.

Il circuito di un raddrizzatore a doppia semionda si compone come segue:

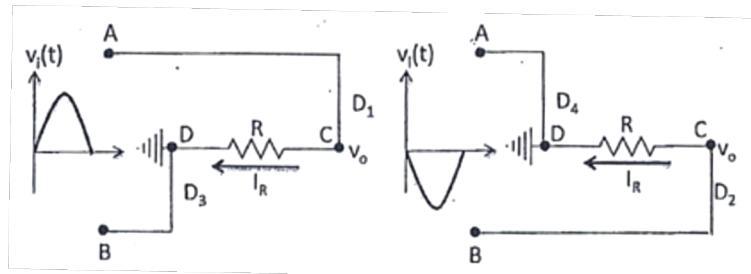


Quando dalla rete proviene la prima semionda si può osservare che il diodo D_4 è interdetto, mentre D_1 è in polarizzazione diretta. Quindi, tutta la tensione positiva di alimentazione si ritrova applicata sul punto C, così che anche D_2 è interdetto e D_3 polarizzato direttamente. Sostituendo i diodi con i cortocircuiti e i circuiti aperti appositi, è possibile misurare una corrente che scorre da C a D e una tensione di uscita $v_o(t)$ sul resistore positiva, che coincide con la semionda $v_i(t)$ in ingresso dalla rete. Quando in ingresso si trova la semionda negativa, per un discorso duale si può dire che D_1 e D_3 sono interdetti e che D_4 e D_2 in polarizzazione diretta. Segue che la tensione negativa di alimentazione si ritrova applicata sul punto D; tuttavia, la corrente che scorre nel carico R sarà ancora diretta da C a D, visto che la tensione $v_o(t)$ è ancora positiva su C rispetto a D e non può che essere uguale a $-v_i(t)$.

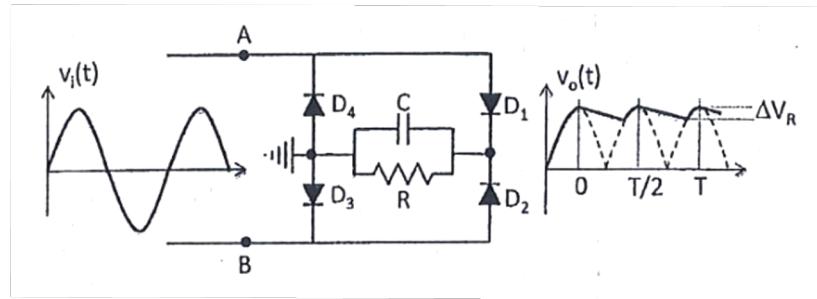
A partire da una semionda positiva e una negativa, il circuito restituisce sul carico due semionde positive, e così via per tutta la lunghezza della sinusoide in ingresso.



Tra una semionda e l'altra si può apprezzare il seguente schema semplificato, composto dallo stesso circuito al quale sono stati sostituiti i cortocircuiti in corrispondenza di diodi in polarizzazione diretta e circuiti aperti in corrispondenza di diodi interdetti:



Se poi si collegasse nel raddrizzatore appena mostrato un condensatore in parallelo al carico la forma d'onda di uscita sarebbe simile a quella del rilevatore di picco ma con un ripple sulla tensione di uscita notevolmente ridotto.



Per il **calcolo del ripple**, si consideri V_P la tensione di picco e $t = 0$ il tempo in cui la tensione di ingresso raggiunge V_P ; l'evoluzione della tensione di uscita è:

$$v_o(t) = V_P e^{-\frac{t}{RC}} \approx V_P \left(1 - \frac{t}{RC}\right)$$

Avendo supposto $RC \gg T/2$, che è la **massima durata della scarica possibile**. Immaginando, per semplicità, che la scarica si interrompa proprio a $T/2$, si ottiene:

$$v_o\left(\frac{T}{2}\right) \approx V_P \left(1 - \frac{T}{2RC}\right)$$

Per cui la **tensione di ripple**, definita come l'incremento tra la fine della scarica e la tensione di picco, è:

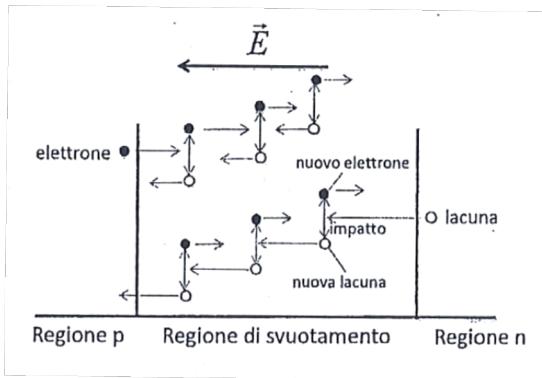
$$\Delta V_r = V_P - v_o\left(\frac{T}{2}\right) = V_P - V_P \left(1 - \frac{T}{2RC}\right) = \frac{V_P T}{2RC} \propto \frac{1}{f}$$

Ed è una **relazione direttamente proporzionale al periodo della forma d'onda in ingresso e inversamente proporzionale alla relativa frequenza**.

LA REGIONE DI BREAKDOWN E IL DIODO ZENER

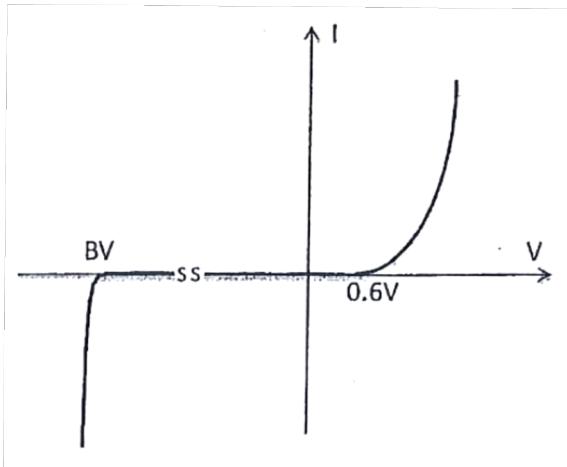
Finora è sempre stato detto che la corrente di polarizzazione inversa di un diodo è ingegneristicamente nulla e questo risultato è stato attribuito alla dipendenza di tale corrente dai portatori minoritari presenti sui due lati della giunzione PN, quantità quasi sempre molto piccola. In realtà, per tensioni inverse molto grandi si possono osservare dei fenomeni per i quali la quantità di portatori minoritari aumenta notevolmente, portando ad un **incremento ripido della corrente**.

Dal punto di vista microscopico, gli elevati valori misurabili di corrente sono da attribuire al **fenomeno della moltiplicazione a valanga**; seguendo il percorso di un portatore minoritario, ad esempio una lacuna, nel suo attraversamento della regione di svuotamento, si può osservare come si muova dalla regione di tipo N alla regione di tipo P con una velocità dipendente dal campo elettrico che innesca il movimento stesso. Durante l'attraversamento della regione di svuotamento, esiste una probabilità non nulla che questi portatori impattino contro gli atomi di silicio e, se possiedono energia cinetica sufficiente, rompino un legame covalente, liberando una coppia elettrone – lacuna. Questa probabilità aumenta di più quanto più è elevata la tensione inversa applicata ai capi della giunzione, aumentando l'accelerazione dei portatori e la loro energia cinetica.



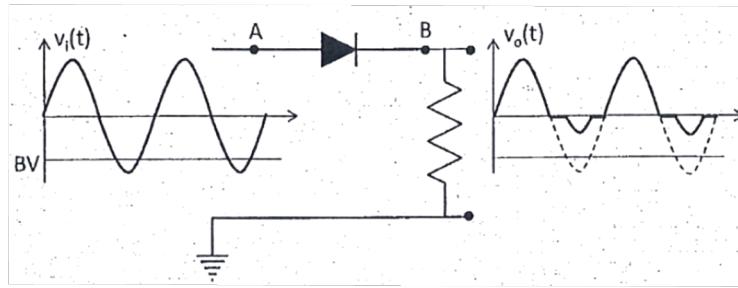
Questo meccanismo è rigenerativo, nel senso che, **una volta innescato**, il numero di cariche che si creano tende ad **aumentare senza limite come un effetto domino**; infatti, **gli elettroni e le lacune liberi che si sono creati dall'impatto sono anch'essi nella regione di svuotamento e sono anch'essi possibili fonti di un ulteriore impatto**. Aumentando a dismisura il numero di portatori minoritari, aumenta a dismisura anche la corrente che circola nel diodo.

Fatte queste considerazioni, **andrebbe aggiornato il modello che descrive un diodo**; analiticamente risulterebbe troppo complesso ma si può **graficare la caratteristica nel piano tensione – corrente (IV)** come segue:

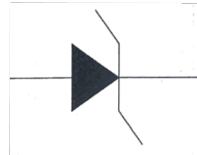


Indicando con $-BV$ la **tensione di breakdown**, ovvero quella **tensione per la quale si innesca il fenomeno della moltiplicazione a valanga** e per il quale si verifica l'aumento ripido della corrente; se nel circuito non sono previsti meccanismi di limitazione della corrente, questo effetto può essere **estremamente deleterio**, essendo la potenza (e quindi il calore) dissipato per effetto Joule quadratico rispetto alla corrente. Il simbolo **SS sul grafico sta ad indicare che ci si è allontanati di molto sulla scala delle tensioni prima di giungere alla $-BV$** , che di per sé è un valore molto elevato (in modulo).

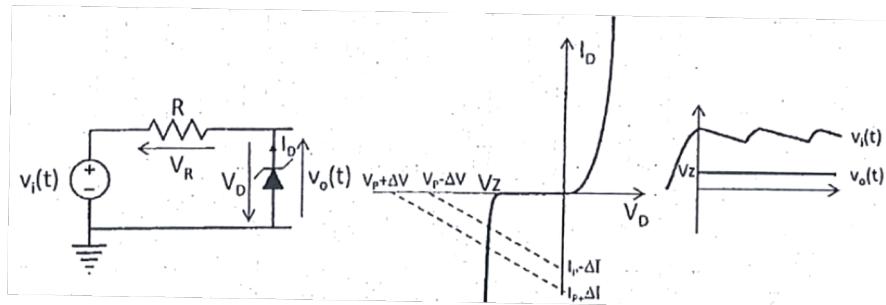
Nel progetto di un circuito con componentistica elettronica nel quale figurano dei diodi va **sempre tenuto in conto il massimo valore di tensione che può essere applicato ai capi dei diodi** in questione; sia fatto l'esempio con il **raddrizzatore a singola semionda** precedentemente analizzato, si nota come quando in ingresso è posta la semionda negativa, al diodo è applicata una **tensione inversa il cui valore massimo è $-V_P$** , che può tranquillamente **superare il valore $-BV$** e corrompere il corretto funzionamento del circuito. Infatti, in questa eventualità, **una buona porzione di semionda negativa della forma di ingresso verrebbe riproposta in uscita**:



La moltiplicazione a valanga necessita di due condizioni per potersi verificare: in primis deve esserci un campo elettrico sufficiente ad una adeguata accelerazione e poi un percorso sufficientemente lungo affinché i portatori possano acquisire una elevata energia cinetica. Può, però, accadere che il campo elettrico sia abbastanza elevato ma il percorso non sufficientemente lungo; in tal caso i legami covalenti possono rompersi non per effetto dell'urto con i portatori ma a causa del campo elettrico stesso, creando una popolazione aggiuntiva di cariche elettriche che conduce ad un brusco aumento della corrente inversa. Questo fenomeno prende il nome di **effetto Zener** e i diodi in cui il breakdown appena descritto avviene per tale effetto a tensioni controllate prendono il nome di **diodi Zener**:



La forma della caratteristica di questo diodo è simile ad uno classico, però con una **tensione di breakdown** (detta anche **tensione di Zener** in questo caso) in genere più bassa (intorno ai 30V), non dipendente dalla temperatura e ingegnerizzata ad hoc per gli utilizzi pratici. La tensione in questione è **estremamente stabile**, tanto che questi diodi vengono usati come **riferimenti di tensione** e negli alimentatori stabilizzati; questi ultimi sfruttano il diodo Zener per appiattire ulteriormente una tensione affetta da **ripple**, magari prodotta da un raddrizzatore, producendo una tensione sostanzialmente costante. Il motivo risiede nel fatto che **nella regione di breakdown la tensione ai capi del diodo è praticamente costante ed indipendente dalla corrente che lo attraversa**.



La tensione $v_i(t)$ è quella prodotta da un raddrizzatore, mentre $v_o(t)$ quella prodotta dal raddrizzamento con diodo Zener. Istante per istante, il bilancio delle tensioni per il circuito è scritto come segue:

$$v_i(t) = V_R - V_D$$

Esprimendo la **caduta di tensione sulla resistenza rispetto alla corrente I_D del diodo Zener** si ha:

$$v_i(t) = -RI_D - V_D$$

Sul piano delle caratteristiche, questa equazione rappresenta una retta che interseca l'asse x in $V_D = -v_i(t)$ e l'asse y in $I_D = -v_i(t)/R$; poiché $v_i(t)$ varia nel tempo, la retta di carico da prendere in considerazione non è sempre la stessa, sebbene possa essere individuata in un intorno di V_P (in $[V_P - \Delta V_r; V_P + \Delta V_r]$) a causa del ripple periodico. La figura precedente mostra chiaramente che, pur variando il punto di intersezione con la caratteristica del diodo, la tensione ai suoi capi (alias tensione di uscita) rimane costante e pari a $v_o(t) = V_Z$; in realtà, la tensione non sarà mai perfettamente costante, ci sarà sempre un ripple ma sarà ingegneristicamente irrilevante. Ovviamente, nella progettazione del circuito è necessario modulare la tensione di ingresso e la resistenza in modo tale che le intersezioni possibili avvengano sempre nella regione di breakdown.

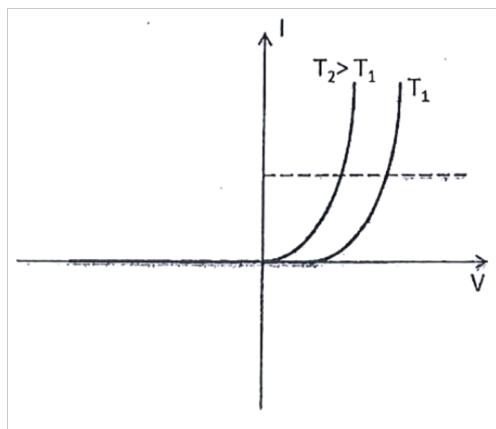
I parametri che caratterizzano il funzionamento dei semiconduttori dipendono fortemente dalla temperatura; quindi, anche le caratteristiche dei dispositivi elettronici realizzati con materiali semiconduttori saranno sensibili alla temperatura. Ad esempio, la caratteristica di un diodo dipende dalla tensione termica, un parametro caratteristico del silicio è definito come:

$$V_t = \frac{kT}{q}$$

Dipendente dalla temperatura. In polarizzazione diretta si può dimostrare che, una volta fissata la corrente, si ha:

$$\frac{dV}{dT} \approx -2.5 \frac{mV}{K}$$

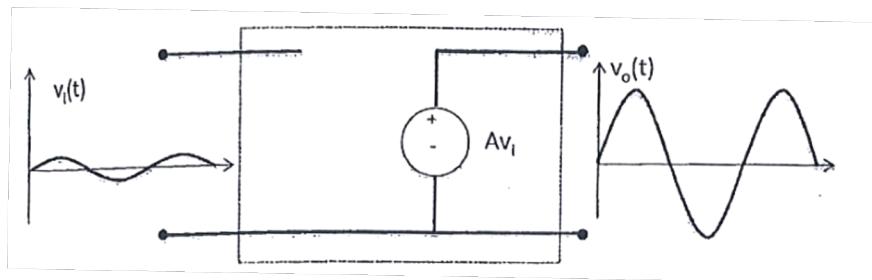
All'aumentare della temperatura, la tensione diretta che bisogna applicare al diodo per ottenere una determinata corrente diminuisce, oppure fissata la tensione aumenta la corrente; sperimentalmente, la corrente inversa in un diodo al silicio raddoppia per ogni 10 gradi di aumento della temperatura.



IL TRANSISTORE BIPOLARE A GIUNZIONE

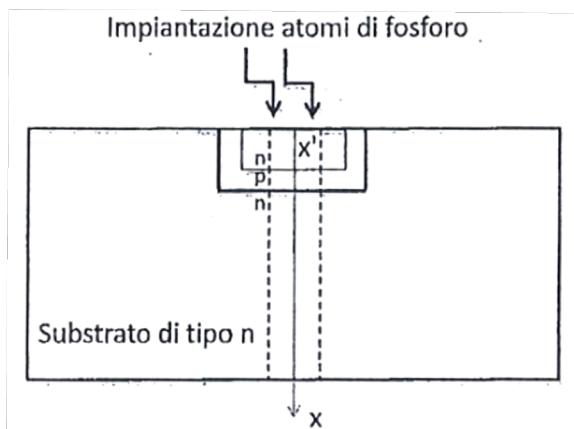
Sin dai principi della trattazione è stata posta particolare attenzione sul processo di amplificazione di un segnale; questa operazione ha lo scopo di prendere un segnale debole (come quelli trasmessi radiofonicamente per limitare l'inquinamento elettromagnetico) e riprodurne una versione fedele

ma con ampiezza maggiore. Il modello generale di **generatore controllato** è, quindi, **un circuito che registra in ingresso la forma del segnale da replicare e la riproduce moltiplicata per una costante in un luogo diverso da quello di prelievo**; un dispositivo di questo genere fa uso di generatori controllati, cioè di componenti che forniscono in uscita una grandezza elettrica che non dipende dallo stato della parte di circuito in cui si trova ma da una grandezza elettrica relativa ad un'altra parte del circuito:

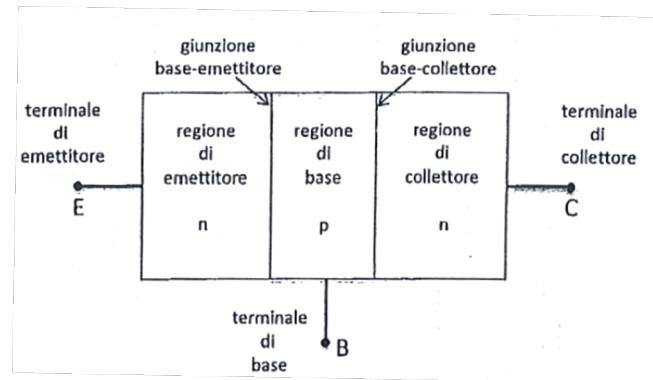


È importante sottolineare che **il generatore controllato modella un fenomeno fisico e che non è un dispositivo fisicamente realizzabile** (soprattutto con bipoli lineari); infatti, **non esiste alcun componente tale che la corrente che lo attraversa non dipenda dalla tensione ad esso stesso applicata**. Per ricercare un comportamento simile si deve andare nel dominio dei dispositivi non lineari, in particolare far uso dei transistori bipolari a giunzione (Bipolar Junction Transistor, BJT).

Il funzionamento di un **transistore bipolare a giunzione** parte dalla **giunzione PN**: è stato precedentemente mostrato come in polarizzazione inversa non circolasse una corrente rilevante; il motivo risiede nel fatto che **il numero di portatori minoritari è decisamente inferiore e che il campo elettrico è tale da favorire solo il passaggio di queste**. Per **aumentare la corrente** che circola in polarizzazione inversa è necessario **aumentare il numero di portatori**; un modo per ottenere questo risultato è la realizzazione **in adiacenza a tale giunzione PN di un'altra giunzione PN**, che viene poi **polarizzata direttamente**:

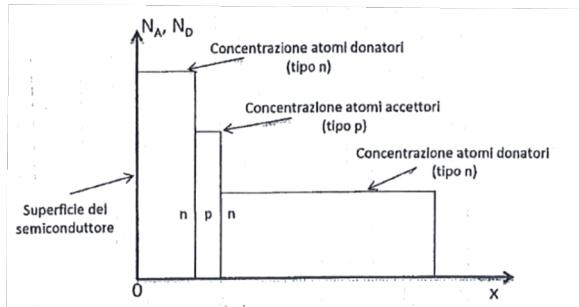


Il dispositivo appena realizzato sarà caratterizzato da **tre regioni distinte, con droggaggio a segni alterni**, chiamate **emettitore, base e collettore**, ognuno dei quali caratterizzato da un proprio terminale:



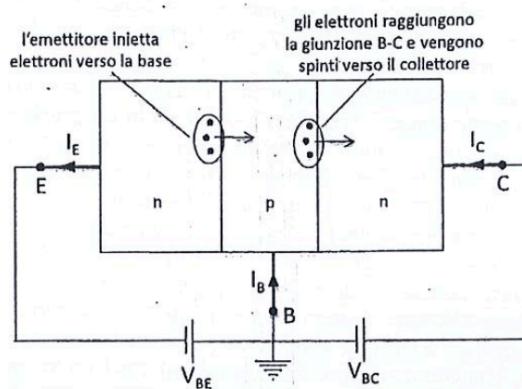
Per motivi che verranno chiariti successivamente, è necessario che il drogaggio di ognuna delle tre regioni segui dei criteri ben precisi; in particolare:

$$N_E \gg N_B \gg N_C$$



Sapendo che N_E e N_C sono le concentrazioni del drogaggio N sull'emettitore e sul collettore e N_B la concentrazione del drogaggio P sulla base.

Tramite gli appositi terminali, si applichi tra la base e il collettore una tensione tale da polarizzare inversamente la relativa giunzione PN (quindi $V_{CB} > 0$): il campo elettrico che si andrà ad instaurare ai capi della giunzione imporrà agli elettroni di migrare dalla base al collettore in una quantità dipendente dalla polarizzazione dell'altra giunzione (tra emettore e base) di cui è composto il dispositivo; in particolare, per tensioni V_{BE} tali da polarizzare direttamente la giunzione, l'emettore (nel quale il numero di elettroni è più elevato) inietta nella base una gran quantità di portatori minoritari, i quali si aggiungeranno ai pochi presenti sulla base per migrare verso il collettore. In queste particolari condizioni si verifica un passaggio di corrente in condizioni di polarizzazione inversa che in un comune diodo non ci sarebbe.



I nomi delle tre regioni non sono scelti a caso: **l'emettitore è quella porzione di semiconduttore che “emette elettroni”, che si appoggeranno sulla base prima di essere “collezionati” dal collettore.** L'effetto fisico su cui preme porre attenzione è quello per il quale la corrente che attraversa il collettore dipende dalla tensione applicata sull'emettitore, non collegato al collettore, ripresentando il comportamento modellato con il generatore controllato. Si può, quindi, pensare che **il transistore permetta di effettuare l'operazione di amplificazione tanto menzionata.**

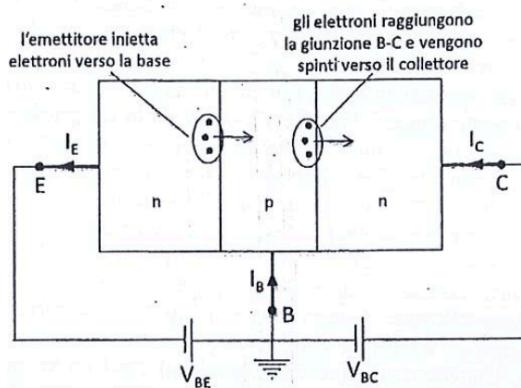
Si noti che, essendo esponenziale il legame tra la tensione applicata ai capi di una giunzione PN e la corrente che vi circola, è esponenziale anche il legame tra la tensione in questione e il numero di portatori minoritari che migra; quindi, la corrente che scorre nel collettore è esponenzialmente dipendente dalla tensione ai capi della giunzione emettitore – base e piccoli incrementi di questa conducono a grandi variazioni di corrente. In altri termini, **il circuito emettitore – base è un circuito a bassa resistenza.**

La corrente di elettroni viene forzata (dal campo elettrico presente sulla giunzione base – collettore) **a fluire verso il collettore indipendentemente dal carico a cui esso è collegato**, potendovi collegare un'alta resistenza e comunque vedere alte correnti nel collettore forzate da basse tensioni all'emettitore. Da queste considerazioni segue che **il transistor è un dispositivo in grado di trasferire la corrente circolante in un circuito a bassa impedenza verso un circuito ad alta impedenza**, e per questo viene chiamato **transfertore di resistenza** (Transfer Resistor → Transistor).

Riassumendo: **il transistor è un dispositivo composto da due giunzioni PN contigue che individuano tre regioni, emettitore – base – collettore; la giunzione base – collettore è polarizzata inversamente per raccogliere sul collettore i portatori minoritari; la giunzione emettitore – base è polarizzata direttamente per permettere la migrazione di una gran quantità di portatori minoritari verso la base, e quindi il collettore; l'idea di controllare la corrente circolante nel collettore attraverso la tensione applicata all'emettitore prende il nome di effetto transistore.**

LE CARATTERISTICHE E LE REGIONI DI FUNZIONAMENTO DI UN BJT

Anche per il BJT, come per il diodo, non verranno ricavate le più rigorose leggi che ne descrivono il comportamento ma **ci si affiderà ad un compromesso tra semplicità e accuratezza.**



Essendo **la giunzione emettitore – base una comune giunzione PN**, la relazione che lega la corrente di emettitore I_E alla tensione applicata ai capi V_{BE} è del tipo:

$$I_E = I_{E0} e^{V_{BE}/V_t}$$

Però questa corrente è composta sia da lacune, che si muovono dalla base all'emettitore, che da elettroni, che fanno il percorso inverso; il motivo per cui $N_E \gg N_B$ è dovuto alla necessità di avere un numero di portatori iniettati dall'emettitore preponderante rispetto al numero di portatori che vi giungono dalla base. Il rapporto tra queste due componenti di corrente viene definito **rendimento di emettitore, γ** , ed è prossimo all'unità nei transistori reali (la corrente di emettitore è composta quasi interamente da elettroni iniettati dall'emettitore). A rigore, **della corrente di collettore fanno parte anche le lacune iniettate dal collettore alla base**, il cui numero è legato alla normale popolazione di portatori minoritari nel collettore, quindi tale corrente è trascurabile.

$$\gamma = \frac{I_{E_n}}{I_{E_n} + I_{E_p}} \approx 1$$

Una volta giunti nella base, gli elettroni diffondono verso il collettore percorrendo mediamente una distanza pari alla lunghezza di diffusione; nella pratica, i transistori vengono realizzati in modo tale da rendere la larghezza della base inferiore alla lunghezza di diffusione, con lo scopo di permettere a più elettroni possibili di arrivare alla giunzione di collettore. Il rapporto tra numero di elettroni che parte dall'emettitore e quello che arriva al collettore è detto **coefficiente di trasporto, α_t** , e in genere è maggiore di 0.9.

Complessivamente, quindi, la corrente di collettore è formata dall'aliquota α_t di corrente di elettroni che parte dall'emettitore, attraversa la base e giunge al collettore; a sua volta, la corrente di elettroni in questione è pari ad un'aliquota γ della corrente totale di emettitore. In definitiva, la corrente di collettore è pari ad un'aliquota $\gamma\alpha_t$ della corrente di emettitore:

$$I_C = \gamma\alpha_t I_E = \alpha I_E = \alpha I_{E0} e^{\frac{V_{BE}}{V_t}} = I_S e^{\frac{V_{BE}}{V_t}}$$

α prende il nome di guadagno di corrente a base comune. L'equazione in questione è la chiave del funzionamento di un BJT; descrive, infatti, la proprietà per la quale la corrente (nel collettore) è controllata da una tensione (ai capi dell'emettitore – base) alla quale non è collegata ed è indipendente dalla tensione (ai capi della base – collettore) applicata alla maglia a cui, invece, è collegata. Resta da trovare una forma per la corrente di base, ottenuta a partire dalla LKC:

$$I_B = I_E - I_C = \frac{1}{\alpha} I_C - I_C = \left(\frac{1-\alpha}{\alpha}\right) I_C = \frac{1}{\beta} I_C$$

Ponendo:

$$\beta = \frac{\alpha}{1-\alpha}$$

Essendo α un valore numerico molto stabile e molto prossimo ad 1, β sarà un valore numerico instabile oscillante e molto grande. Il parametro β rappresenta il principale parametro di merito dei transistori bipolarì a giunzione e prende il nome di **guadagno di corrente ad emettitore comune**.

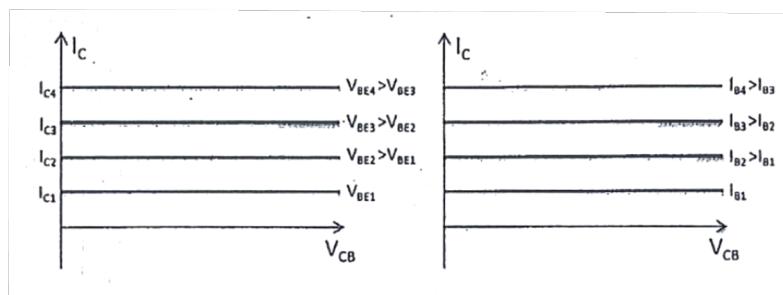
Riassumendo le equazioni che descrivono le correnti nel BJT:

$$\begin{cases} I_C = I_S e^{\frac{V_{BE}}{V_t}} \\ I_B = \frac{1}{\beta} I_C \\ I_E = \frac{1}{\alpha} I_C \end{cases}$$

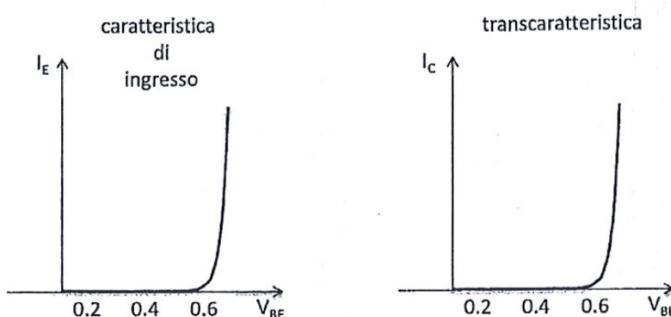
Osservando queste tre correnti per valori reali di α e β , si nota che I_C è sostanzialmente uguale a I_E e che I_B è molto piccola, tanto da poter essere spesso trascurata.

La rappresentazione grafica tensione – corrente (IV) di questi modelli per il BJT non è univoca; infatti, dipende da quali grandezze si prendono in considerazione, essendoci diverse correnti e diverse tensioni coinvolte. Si parla di **caratteristica di uscita quando le grandezze prese in considerazione sono quelle relative al circuito di collettore, I_C e V_{BC}** ; in realtà, essendo la rappresentazione fatta finora a base comune, si può dire che questa caratteristica è di **uscita a base comune**, mentre è di **ingresso a base comune se è relativa al circuito di emettitore, con I_E e V_{BE}** . Infine, si parla di **transcaratteristica quando le grandezze usate per la rappresentazione non sono legate nella stessa maglia** (ad esempio, correnti di uscita e tensioni di ingresso, e così via...).

Le caratteristiche di interesse pratico sono quelle di uscita, quindi con la tensione V_{BC} (che si suppone negativa per rientrare nel dominio di funzionamento precedentemente mostrato) e con la corrente I_C (che è la grandezza di uscita vera e propria). Dalle relazioni rilevate, si può intuire che la corrente in esame I_C non è dipendente dalla relativa tensione V_{BC} , bensì dalla tensione in ingresso V_{BE} ; di conseguenza, sul piano tensione – corrente (IV) la caratteristica di uscita sarà rappresentata da un fascio di rette parallele all'asse delle tensioni, le cui distanze l'una dalle altre dipende unicamente dalla tensione V_{BE} a cui ognuna è associata. Le rette in questione non sono tutte equidistanti, dal momento in cui la relazione tra I_C e V_{BE} è esponenziale:

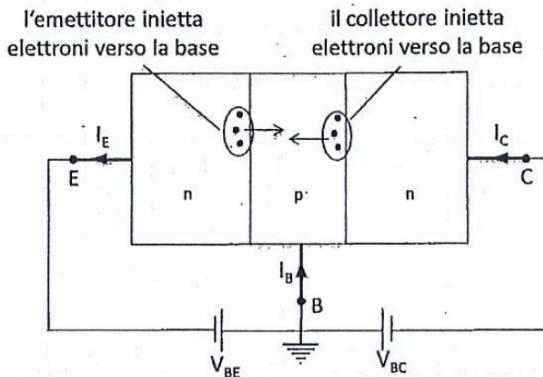


Per quanto riguarda la **caratteristica di ingresso**, la relazione tra I_E e V_{BE} è un semplice esponenziale che è stato visto anche per la giunzione PN. Una forma simile, ovviamente, si presenta anche per la **transcaratteristica $I_C - V_{BE}$** :



Le caratteristiche finora osservate si limitano al primo quadrante (in particolar modo al semiasse positivo delle ascisse), **nonostante ai capi delle giunzioni PN possa essere applicata qualsiasi tensione**. Il motivo risiede nel fatto che quella che si è analizzata, per cui $V_{BE} > 0$ e $V_{BC} < 0$, è la **regione di funzionamento che innesca l'effetto transistore**, detta anche **Regione Attiva Diretta** (RAD).

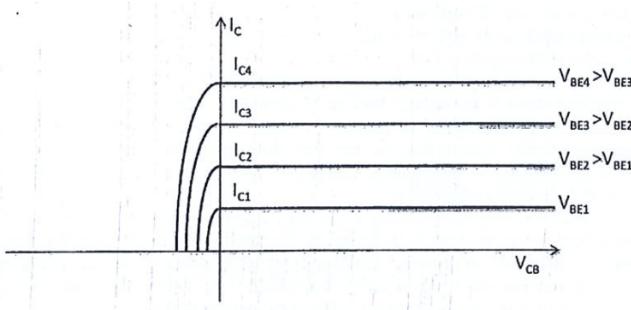
Prima di osservare la caratteristica di uscita ampliata, **si analizzi il funzionamento del transistore nel caso in cui $V_{BC} > 0$ (regione di saturazione)**:



Il campo elettrico presente sulla giunzione emettitore – base non cambia segno, quindi **non cambia nulla nei riguardi della corrente di elettroni proveniente dall'emettitore**, tali elementi vengono comunque spinti verso il collettore; tuttavia, **in polarizzazione diretta il collettore inietta anch'esso elettroni nella base**, i quali si comportano simmetricamente a quelli iniettati dall'emettitore. La corrente totale che circola nel dispositivo è legata ai due flussi indipendenti di elettroni, quindi alla differenza delle due relative correnti che sono rappresentate da una legge esponenziale:

$$I_C = I_{E0} e^{\frac{V_{BE}}{V_t}} - I_{C0} e^{\frac{V_{BC}}{V_t}}$$

Ed è una **quantità prossima allo zero perché i due contributi tendono ad annullarsi**. Per ogni fissato valore di V_{BE} , quando la tensione V_{BC} è positiva la corrente di collettore inizia a diminuire ed esiste sicuramente un valore (attorno agli 0.8V perché comunque è una giunzione PN polarizzata direttamente) per cui $I_C = 0$. **Graficamente**, questa relazione è rappresentata come segue:



La regione di funzionamento che corrisponde alle due giunzioni polarizzate direttamente è detta **regione di saturazione**; un'altra condizione di funzionamento prevede $V_{BC} < 0$ e $V_{BE} < 0$, quindi $I_C \approx I_E \approx 0$, ed è detta **regione di interdizione**. L'ultima condizione di funzionamento prevede che collettore ed emettitore si scambino di ruolo, quindi $V_{BE} < 0$ e $V_{BC} > 0$. Nel caso in cui le due regioni fossero tecnologicamente identiche, il transistore funzionerebbe specularmente a come fa

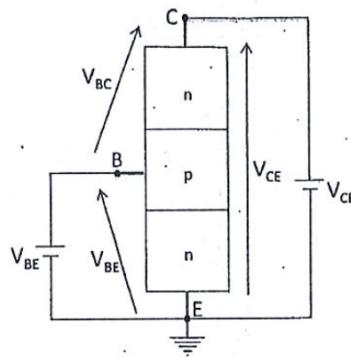
in RAD; tuttavia, essendo l'emettitore drogato molto di più rispetto al collettore, in queste condizioni di funzionamento il dispositivo ha dei parametri molto poveri (si parla di $\alpha \ll 1$ e β molto elevato) che rendono il transistore inutilizzabile. La regione di funzionamento in questione è detta **Regione Attiva Inversa (RAI)**.

Riassumendo:

	$V_{BE} > 0$	$V_{BE} < 0$
$V_{BC} > 0$	Saturazione	RAI
$V_{BC} < 0$	RAD	Interdizione

RAPPRESENTAZIONE AD EMETTITORE COMUNE E L'EFFETTO EARLY

In molte applicazioni pratiche il terminale di riferimento delle tensioni non è la base ma l'emettitore, dove la tensione al collettore viene applicata tramite il generatore V_{CE} . Il funzionamento del dispositivo non cambia in questa configurazione, sebbene sia necessario capire, per ogni V_{CE} applicata, quale sia la corrispettiva tensione che insiste sulla giunzione base – collettore determinandone la polarizzazione inversa o diretta. Sulla base di quanto appena detto e facendo riferimento alla seguente figura, si provino a tracciare le caratteristiche di uscita ad emettitore comune (I_C in funzione di V_{CE} prendendo V_{BE} come parametro).



Si può chiaramente intuire che:

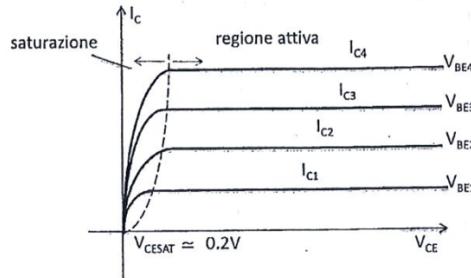
$$V_{CB} = V_{CE} - V_{BE}$$

Ovvero che, assegnato V_{BE} , la giunzione base – collettore è polarizzata inversamente se $V_{CE} > V_{BE}$; si lavora, quindi, in RAD e si può usare l'equazione:

$$I_C = I_S e^{\frac{V_{BE}}{V_t}}$$

Graficamente significa che la curva caratteristica corrispondente a ogni V_{BE} si mantiene costante per $V_{CE} > V_{BE}$ ed è figurata come la caratteristica precedentemente individuata ma traslata di V_{BE} . Quando, invece, $V_{CE} < V_{BE}$, la giunzione base – collettore è polarizzata direttamente, il

transistore entra nella regione di saturazione e la corrente di collettore tende a zero, come era già stato individuato precedentemente.

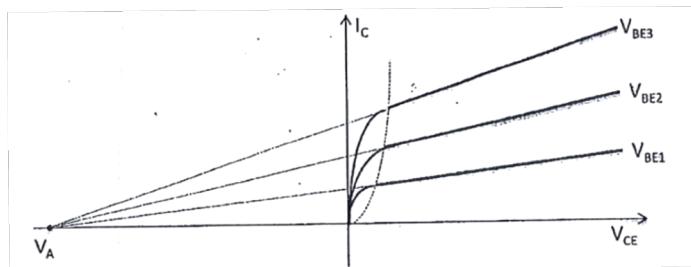


Si può pensare che **ogni curva passi per l'origine** ma ciò sarebbe vero solo se il transistore fosse **simmetrico**, ovvero se $I_{C0} = I_{E0}$; in realtà, essendo l'emettitore più drogato del collettore, in scala espansa si vedrebbero le varie curve intersecare l'asse delle ascisse per V_{CE} leggermente positivo (pochi millivolt).

Quando le giunzioni sono entrambe polarizzate direttamente, si ha $V_{CE} = V_{BE} - V_{BC}$, con entrambi gli addendi inferiori a 0.8V; ciò significa che, in ogni caso, la **regione di saturazione del transistore si estende per non più di $V_{CE} = 0.8V$** . Tenendo anche conto che la giunzione base – collettore (come tutte le giunzioni) non inietta significativamente fino a che la tensione ai suoi capi non supera la soglia di 0.6V, si rileva che la **tensione che si trova ai capi del transistore quando questo è in saturazione è $V_{CE_{sat}} = 0.2V$** . Quando ci si trova in regione di saturazione si usa convenzionalmente questo valore.

Le caratteristiche di uscita del transistore vengono mostrate indipendenti rispetto alla tensione che l'alimentazione V_{CE} impone sulla giunzione base – collettore; nella realtà il legame non è di perfetta indipendenza ma di una lievissima dipendenza lineare che prende il nome di effetto Early. Per ottenere una modellizzazione dell'effetto Early si deve considerare il punto di convergenza sull'asse x di tutte le caratteristiche al variare di V_{BE} , quindi il punto V_A (detta tensione di Early) per il quale:

$$I_C = I_S e^{\frac{V_{BE}}{V_t}} \left(1 + \frac{V_{CE}}{V_A} \right)$$

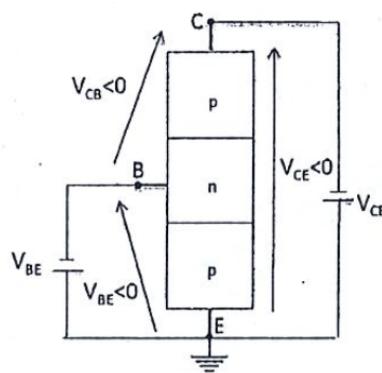


C'è, quindi, la **dipendenza sia dalla V_{BE} che dalla V_{CE}** ; tuttavia, quest'ultima influenzera poco la forma d'onda perché, generalmente, V_A è un valore molto lontano nel semiasse negativo delle tensioni. I due modelli finora mostrati non sono del tutto scollegati tra di loro: il modello V_{CE} – indipendente può essere considerato il limite per $V_A \rightarrow -\infty$ del modello V_{CE} – dipendente.

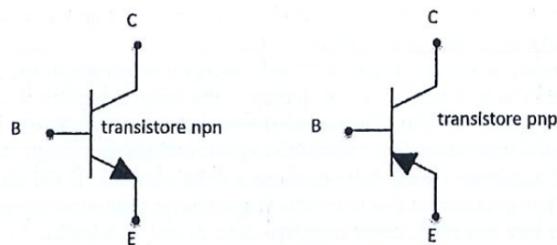
L'effetto Early è una conseguenza della polarizzazione inversa; infatti, in tali condizioni, la regione di svuotamento è più larga, con un parametro α_t più elevato ed una maggior probabilità

che l'elettrone viaggi lungo tutte e tre le giunzioni, restituendo **più corrente** e legando linearmente le due grandezze.

Un transistore bipolare a giunzione può essere creato e usato sia come finora specificato, quindi NPN, sia nella sua versione duale, PNP, in cui la base è drogata di tipo N e collettore ed emettitore di tipo P. Il funzionamento dei due dispositivi è identico, a patto di cambiare opportunamente i segni di tensioni e correnti; l'unica differenza significativa è nelle correnti: essendo queste dipendenti dalla mobilità dei portatori, ad un transistore PNP sarà associata la mobilità delle lacune, minore di circa 2.5/3 volte di quella degli elettroni, associata al transistore NPN. Di conseguenza, un transistore PNP ammetterà una corrente inferiore ad un transistore NPN, sebbene si comportino allo stesso modo.

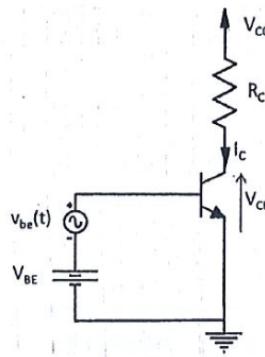


Per individuare un transistore si utilizza il **simbolo circuitale seguente, caratterizzando la base dalla linea verticale, l'emettitore dalla freccia e il collettore dall'arco obliquo rimanente. Il verso della freccia sta ad indicare il verso della giunzione PN base – emettitore (quasi come fosse un diodo): per un BJT NPN punterà lontano dalla base, per un BJT PNP punterà verso la base. Infine, se si volessero associare le correnti sui tre terminali, per un BJT NPN esse saranno tutte entranti, per un BJT PNP tutte uscenti.**



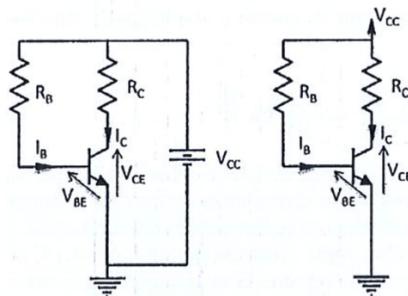
POLARIZZAZIONE DI UN BJT E CIRCUITI AMPLIFICATORI

Una delle applicazioni pratiche dei BJT più comuni è quella di amplificatore; l'amplificazione di un segnale da parte di un BJT passa per un processo di polarizzazione che ha lo scopo di stabilizzare le tensioni che consentono al transistore di lavorare in specifiche regioni di funzionamento. Il modello più astratto di amplificatore è il seguente:



Tuttavia, si scorgono già adesso delle problematiche non di poco conto: in primis, **il circuito suppone due alimentazioni** (si sorvola il generatore di segnale perché non è un'alimentazione ma una forma d'onda in ingresso al circuito), ma è anche **troppo sensibile alle temperature**; infatti, le **caratteristiche di funzionamento di un BJT sono sensibili ai parametri α e β** , che a loro volta dipendono dalla temperatura (se la temperatura aumenta, aumenta leggermente α ma aumenta notevolmente β , e di conseguenza il punto di funzionamento cambia). Lo scopo della polarizzazione di un dispositivo di questo tipo è di imporre il funzionamento in **Regione Attiva Diretta** (così da poter effettivamente amplificare) e di rendere il tutto meno dipendente possibile dal parametro β (oppure, rendere il circuito resiliente), così che il dispositivo possa necessitare una sola alimentazione (come i comuni cellulari o computer) e possa essere usato anche in ambienti a diverse temperature.

Un primo passo in avanti è il **circuito di polarizzazione a due resistenze** (per questa prima volta è proposto sia lo schema integrato che quello unifilare, che successivamente sarà preferito per una questione di semplicità grafica):



Si vogliono determinare i valori di resistenza R_B e R_C necessari per far circolare nel collettore una corrente assegnata, supponendo noto dal data sheet il parametro β . Si possono individuare già due maglie: una **maglia di ingresso**, che comprende la V_{BE} , e una **maglia di uscita**, che comprende la V_{CE} ; l'equazione alla maglia di ingresso prevede che:

$$V_{CC} = R_B I_B + V_{BE}$$

Per cui:

$$I_B = \frac{V_{CC} - V_{BE}}{R_B}$$

Supponendo $V_{BE} = 0.7V$ per convenzione e con V_{CC} assegnato, si può direttamente ricavare il valore di R_B necessario. Per quanto riguarda la **corrente di collettore**:

$$I_C = \beta I_B = \beta \frac{V_{CC} - V_{BE}}{R_B}$$

Ovviamente, la relazione appena mostrata è valida solo se il transistore lavora in RAD, cioè se $V_{CE} > V_{BE}$; quindi, V_{CE} va scelta arbitrariamente in modo tale da garantire un margine di sicurezza rispetto a V_{BE} (tale margine influenza le prestazioni degli amplificatori).

L'equazione alla maglia di uscita prevede che:

$$V_{CC} = R_C I_C + V_{CE}$$

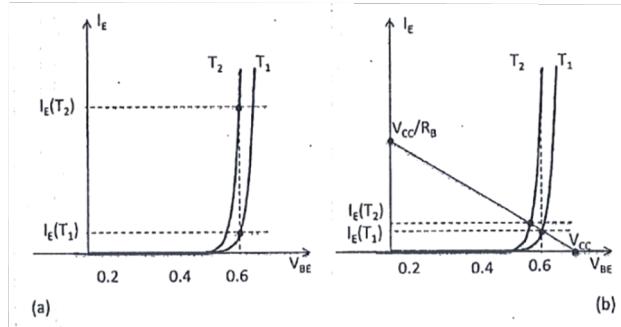
Poiché I_C e V_{CC} sono assegnati e poiché V_{CE} è arbitrariamente scelta (entro i limiti della RAD), l'unica incognita è R_C :

$$R_C = \frac{V_{CC} - V_{CE}}{I_C}$$

Già si possono evidenziare dei punti di forza e delle debolezze di questa polarizzazione:

- **Vantaggi**

La rete resistiva rappresenta un vantaggio rispetto all'applicazione diretta dei generatori di tensione alle giunzioni. Facendo riferimento alla caratteristica di ingresso del transistore ($I_B - V_{BE}$) riportata per due temperature ($T_2 > T_1$, sapendo che la variazione in relazione alla temperatura di un diodo e di un BJT è la stessa):



Si nota che, se tra base ed emettitore viene applicato direttamente un generatore che impone la V_{BE} (ad esempio, a 0.6V), la corrente di base alla temperatura T_1 sarebbe di gran lunga inferiore che alla temperatura T_2 , con analoghe variazioni su I_C . In corrispondenza della stessa variazione di temperatura, la variazione della corrente di base è notevolmente inferiore nel circuito di polarizzazione a due resistenze, come si può anche notare con la retta di carico relativa al circuito in ingresso.

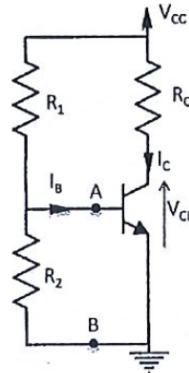
- **Svantaggi**

Questo tipo di polarizzazione non risolve ancora il problema della temperatura e lo si può osservare dalla dipendenza di I_C da β ; i valori nominali di questo parametro, infatti, vanno intesi come indicativi, con range di variazioni che possono superare il 50%. Il motivo di questa variabilità di β risiede nella dipendenza critica da α :

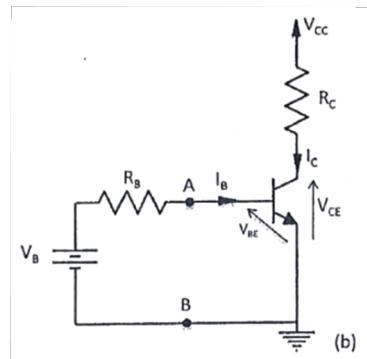
$$\beta = \frac{\alpha}{1 - \alpha}$$

Piccole variazioni di α conducono ad enormi variazioni di β . **Lo svantaggio è ancora preponderante ed il circuito di polarizzazione a due resistenze non è sufficiente per costruire un buon amplificatore di segnale.**

In molti casi, è utile disporre di un grado di libertà in più per dimensionare il circuito di ingresso; la risposta a questa esigenza è il **circuito di polarizzazione a tre resistenze**:



Il partitore resistivo composto da R_1 e da R_2 permette di regolare la tensione applicata alla base come frazione della tensione di alimentazione. Applicando il teorema di Thevenin tra i punti A e B:

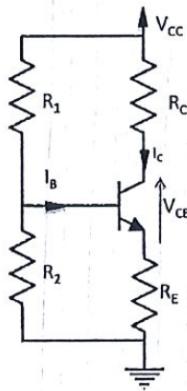


$$V_B = V_{CC} \cdot \frac{R_2}{R_1 + R_2} = V_{CC} \cdot \frac{1}{\frac{R_1}{R_2} + 1}$$

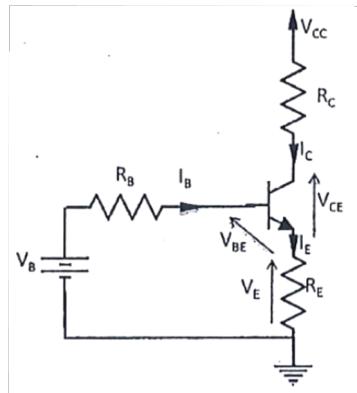
$$R_B = R_1 || R_2 = \frac{R_1 R_2}{R_1 + R_2}$$

Topologicamente, questo circuito è identico a quello di polarizzazione a due resistenze, con la differenza che qui il valore di V_B imposto dal partitore dipende dal rapporto tra R_1 e R_2 , non singolarmente dai due valori di resistenza (che invece influenzano le prestazioni dell'amplificatore, da cui discende l'utilità di questo modello). Il modello appena introdotto non risolve alcun problema precedentemente rilevato, visto che topologicamente non si sta effettuando alcuna modifica.

La polarizzazione di un BJT con quattro resistenze risolve il problema della temperatura; un circuito di questo tipo è composto come segue:



Similarmente al circuito precedente, applicando il teorema di Thevenin:



Indicando con V_E la caduta di tensione ai capi della resistenza R_E , si può scrivere:

$$I_B = \frac{V_B - V_{BE} - V_E}{R_B}$$

Ma, considerando che:

$$V_E = R_E I_E \wedge I_E \approx I_C \wedge I_B = \frac{I_C}{\beta}$$

$$I_C = \frac{V_B - V_{BE}}{R_E + \frac{R_B}{\beta}}$$

Si ricordi che β è un numero molto grande, scegliendo opportuni valori per cui:

$$R_E \gg \frac{R_B}{\beta}$$

Si ha che:

$$I_C \approx \frac{V_B - V_{BE}}{R_E}$$

Ottenendo una relazione pressoché indipendente da β , rendendo la corrente di collettore molto più stabile rispetto ai circuiti di polarizzazione precedenti (soprattutto per variazioni di temperatura).

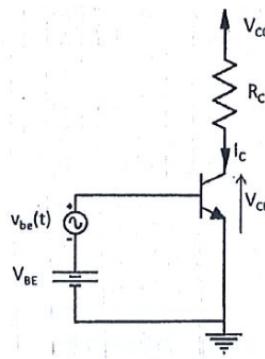
Inoltre, se ci si trova in condizioni per cui $V_B \gg V_{BE}$, si può fare un'ulteriore approssimazione e dire che:

$$I_C \approx \frac{V_B}{R_E}$$

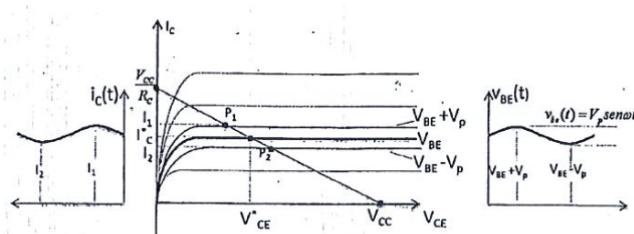
In queste ipotesi, la corrente di collettore assume valori completamente indipendenti dalle caratteristiche fisiche del transistore; si ritrova il comportamento da generatore controllato del transistore, riducendo significativamente l'impatto del dispositivo stesso sul processo di amplificazione. Rendere indipendenti le operazioni che svolgono i dispositivi elettronici dai parametri degli stessi è un procedimento spesso ricercato nell'elettronica, finalizzato ad indurre la minore distorsione possibile

Per quello che riguarda l'elettronica, un segnale è una grandezza fisica variabile nel tempo che, per comodità, verrà rappresentata come somma di sinusoidi con il teorema di Fourier. I circuiti precedenti non vedono in ingresso una forma d'onda variabile nel tempo, sono stati usati solo per determinarne la polarizzazione; determinare preliminarmente il punto di funzionamento è utile per due motivi. In primis, quando si parla di amplificazione si parla del processo di restituzione dello stesso segnale in ingresso con maggiore ampiezza; l'energia necessaria per costruire questo nuovo segnale non può essere né fornita dall'etere gratuitamente né già presente nel segnale di ingresso (l'abbassamento dell'energia è uno dei motivi per cui la trasmissione è fatta su segnali poco ampi) ma è il circuito di polarizzazione che la fornisce tramite l'alimentazione in continua. Va quindi ben tenuto in considerazione che il segnale in ingresso, sebbene sia stato rappresentato come un generatore di tensione, non è una sorgente di alimentazione, bensì una sorgente di informazione.

Per comprendere il secondo motivo bisogna ritornare alla raffigurazione dell'amplificatore nella sua forma generale:



In ingresso al sistema sono posti due generatori di tensione, uno sinusoidale e uno continuo; osservando le caratteristiche di uscita si può notare che la tensione applicata alla maglia di ingresso del transistore non è costante:



Quando varia la tensione in ingresso cambia anche l'intersezione con la retta di carico, conducendo all'amplificazione effettiva del segnale tramite il valore assunto da V_{CE} (che segue le variazioni della tensione di ingresso); tuttavia, se la componente sinusoidale fosse assente, l'intersezione con la retta di carico sarebbe sempre la stessa, dal momento in cui alla maglia di ingresso del transistore non ci sarebbero tensioni variabili e la caratteristica di uscita sarebbe sempre la stessa.

Si può, quindi, intuire che **la polarizzazione del circuito determina il punto di funzionamento a riposo V_{CE}^*** , cioè il punto di funzionamento che si raggiunge quando in ingresso all'amplificatore è posto un segnale nullo.

Si vuole fare una breve digressione sulla **nomenclatura**; in ingresso al transistore è posta la serie tra il generatore costante V_{BE} e quello sinusoidale $v_{be}(t)$:

$$v_{BE}(t) = V_{BE} + v_{be}(t)$$

Con lettera minuscola e pedice maiuscolo si intende **una grandezza variabile sinusoidale**, lettera minuscola e pedice minuscolo il **segnale sinusoidale in ingresso** e lettera maiuscola e pedice maiuscolo **una grandezza costante**. La variazione di $v_{be}(t)$ permette di dire che la tensione in ingresso $v_{BE}(t)$ oscilla tra $V_{BE} + V_p$ e $V_{BE} - V_p$, con V_p valore di picco di $v_{be}(t)$. Si può supporre in maniera analoga che la corrente di collettore vari in maniera sinusoidale intorno al valore I_C^* , assumendo tutti i valori compresi tra I_1 e I_2 ; la corrente complessiva che circola nel collettore è:

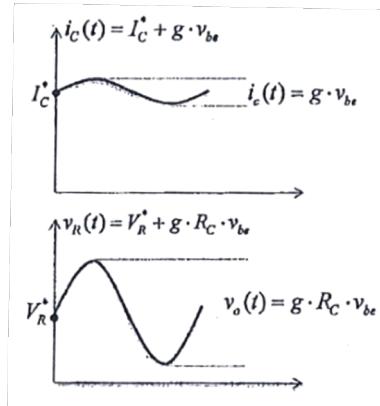
$$i_C(t) = I_C^* + i_c(t)$$

La parte costante è legata al punto di funzionamento a riposo, mentre quella sinusoidale al segnale. È ovvio che l'amplificazione passa per le ampiezze delle componenti sinusoidali della corrente e della tensione:

$$i_c(t) = g \cdot v_{be}(t)$$

Dove **g è la conduttanza di amplificazione**. In conclusione, nel resistore R_C circola una corrente $i_c(t)$ che varia in funzione del segnale attorno a I_C^* :

$$v_R(t) = R_C i_C(t) = R_C I_C^* + R_C i_c(t) = V_R^* + R_C \cdot g \cdot v_{be}(t)$$



Ai capi del resistore R_C è presente una componente di tensione di forma sinusoidale che risulta proporzionale al segnale di ingresso $v_{be}(t)$:

$$v_o(t) = R_C g \cdot v_{be}(t)$$

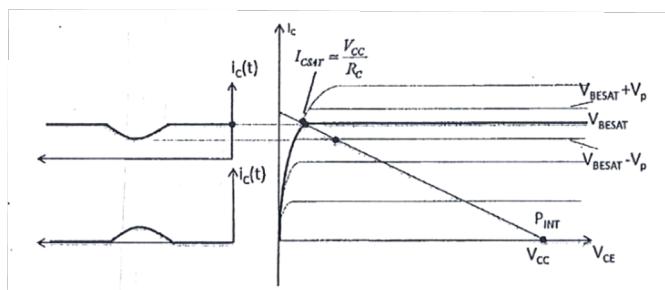
Con costante di amplificazione (o guadagno):

$$A = R_C \cdot g$$

Agendo sul valore di R_C (entro certi limiti), l'amplificatore può essere realizzato con un guadagno arbitrario.

Sulla base di quanto appena detto, è fondamentale la scelta del punto di funzionamento per ottenere una corretta amplificazione; prima di andare a modulare la polarizzazione di un amplificatore è bene chiedersi quali siano i limiti e i vincoli entro cui lavorare.

Il vincolo principale, imposto dal circuito di uscita, è l'appartenenza alla retta di carico; infatti, qualunque sia la V_{BE} applicata in ingresso, il punto di funzionamento deve sempre giacere sulla retta di carico.



È evidente, soprattutto osservando la figura, che quando la V_{BE} aumenta, ci si avvicina sempre di più ad un punto di funzionamento in regione di saturazione. Applicando una V_{BE} sufficiente a far entrare il transistore in saturazione ($V_{BE} = V_{BESat}$), la corrente di collettore non può più crescere, dal momento in cui in quella zona le varie curve a V_{BE} costante sono praticamente sovrapposte:

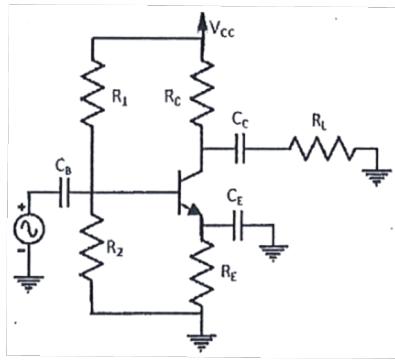
$$I_C = \frac{V_{CC} - V_{CE}}{R_C} \approx \frac{V_{CC}}{R_C}$$

Ovvero, in nessun caso si può avere $I_C > V_{CC}/R_C$ perché la tensione V_{CE} si riduce a decimi di volt trascurabili.

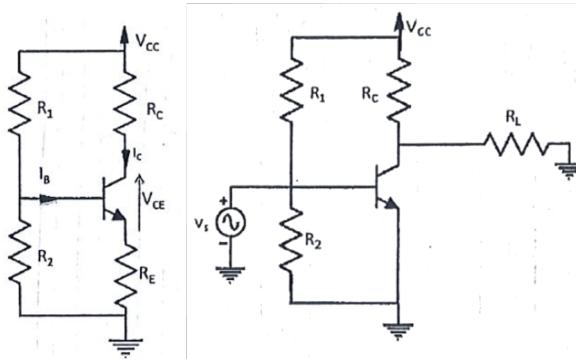
Supponendo di voler scegliere come punto di funzionamento a riposo proprio quello corrispondente a V_{BESat} e I_{Csat} , si individui la forma d'onda ottenuta sovrapponendovi $v_{be}(t)$ con valore di picco V_p . In queste ipotesi, si può facilmente intuire che la corrente di collettore non può seguire le variazioni del segnale di ingresso, visto che quando la tensione $v_{BE}(t)$ sale sopra il valore di V_{BESat} la corrente di collettore resta bloccata a I_{Csat} ; la forma d'onda in uscita è quella mostrata in figura, con una intollerabile perdita di informazione. La regola generale per evitare questi comportamenti prevede che il punto di funzionamento sia scelto in modo che la massima escursione del segnale non porti il transistore a lavorare in regione di saturazione.

Un ragionamento analogo può essere fatto per punti di funzionamento prossimi alla regione di interdizione; quindi, generalmente, il punto di funzionamento a riposo deve essere scelto in modo che l'escursione del segnale in ingresso imponga il transistore a lavorare in RAD, ovvero con valori di V_{CE} statica (V_{CE}^* nelle figure precedenti) sufficientemente maggiori di V_{CESat} ma sufficientemente minori dell'alimentazione V_{CC} .

Si consideri il seguente circuito:



Non è molto differente dal circuito di polarizzazione a quattro resistenze, se non per il **collegamento ad un generatore di segnale, ad un carico e alla massa tramite i tre condensatori C_B , C_C e C_E** . I condensatori sono impiegati per separare, anche fisicamente, il comportamento dinamico dell'amplificatore (AC) da quello statico (DC); infatti, rispetto a grandezze continue i condensatori si comportano come circuiti aperti, garantendo che il punto di funzionamento non sia alterato né dal segnale né dal carico, ma rispetto a frequenze sufficientemente alte del segnale come cortocircuiti, garantendo l'operatività del dispositivo.



IL MODELLO EQUIVALENTE A PICCOLO SEGNALE

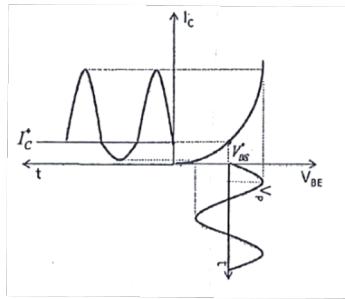
Nel descrivere il funzionamento da amplificatore, è stato ipotizzato che le variazioni della corrente di collettore dipendessero linearmente dalle variazioni del segnale applicato tra base ed emettitore:

$$i_c(t) = g \cdot v_{be}(t)$$

Imponendo la stessa forma d'onda ma con i valori di picco g volte maggiori. Volendo essere precisi, la relazione che lega la corrente di collettore con la tensione tra base ed emettitore è:

$$i_c(t) = I_S e^{\frac{v_{BE}(t)}{V_T}} = I_S e^{\frac{V_{BE} + v_{be}(t)}{V_T}} = I_S e^{\frac{V_{BE}}{V_T}} e^{\frac{v_{be}(t)}{V_T}} = I_C^* e^{\frac{v_{be}(t)}{V_T}}$$

Ed è una **relazione tutt'altro che lineare**; infatti, se la tensione in ingresso è sinusoidale, la corrente di collettore assume la forma di un'esponenziale di una sinusoide. La forma del segnale non si conserva attraversando l'amplificatore, si distorce in funzione dell'ampiezza del segnale variabile; tale effetto è dovuto alla non linearità del transistore e può essere visualizzato nella figura seguente:



Se, però, la variazione del segnale è abbastanza piccola da poter permettere, in un intorno del punto di funzionamento a riposo, l'approssimazione dell'esponenziale alla sua tangente, è possibile osservare nuovamente una **variazione di corrente direttamente proporzionale alle variazioni di tensione**. L'operazione appena eseguita, detta **linearizzazione della caratteristica**, è possibile solo se il segnale di ingresso è piccolo ($x \ll 1$), in modo da poter eseguire l'**espansione in serie di Taylor con un errore ingegneristicamente trascurabile**:

$$e^x \approx 1 + x$$

In termini di tensioni:

$$i_c(t) = I_C^* e^{\frac{v_{be}(t)}{V_t}} \approx I_C^* + I_C^* \cdot \frac{v_{be}(t)}{V_t} = I_C^* + g_m \cdot v_{be}(t)$$

Ovvero, se $v_{be}(t) \ll V_t = 25mV$; questa condizione prende il nome di **condizione di piccolo segnale**. Volendo estendere il concetto di piccolo segnale a tutte le altre grandezze caratteristiche dell'amplificatore:

$$i_B(t) = \frac{I_C^*}{\beta} e^{\frac{v_{be}(t)}{V_t}} \approx I_B^* + \frac{I_B^*}{V_t} v_{be}(t) = I_B^* + \frac{v_{be}(t)}{r_\pi}$$

Con $r_\pi = V_t/I_B^*$. Da queste relazioni si ricava:

$$\beta = g_m \cdot r_\pi$$

Per la corrente di emettitore:

$$i_E(t) = \frac{I_C^*}{\alpha} e^{\frac{v_{be}(t)}{V_t}} \approx I_E^* + \frac{I_E^*}{V_t} v_{be}(t) = I_E^* + \frac{v_{be}(t)}{r_e}$$

Con $r_e = V_t/I_E^*$. È anche facile ricavare che:

$$r_e = \frac{r_\pi}{\beta + 1}$$

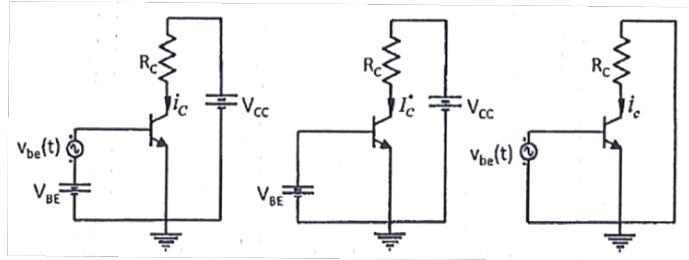
$$g_m \cdot r_e = \alpha$$

Infine:

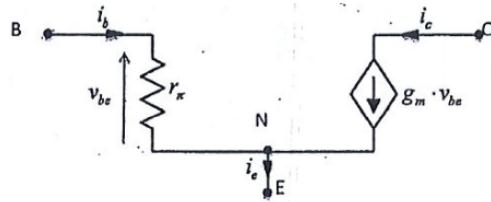
$$v_{CE}(t) = V_{CC} - R_C i_C(t) = V_{CC} - R_C I_C^* - R_C \cdot g_m v_{be}(t) = V_{CE}^* - R_C \cdot g_m v_{be}(t)$$

Si è ottenuto un quadro in cui **tutte le grandezze del circuito sono esprimibili come combinazione lineare di componenti che dipendono**, indipendentemente gli uni dagli altri, **dai singoli generatori**

presenti nel dispositivo, con la conseguente **possibilità di applicare il principio di sovrapposizione degli effetti**. Se si ricorre allo schema ad una resistenza dell'amplificatore, **il principio di sovrapposizione degli effetti** non va applicato ai singoli generatori, bensì **al contributo AC/DC**: si considera **prima la risposta in AC** (quindi con entrambi i generatori di tensione continua accesi), **poi quella in DC** (quindi con entrambi i generatori di tensione continua spenti) e, infine, **si sommano i contributi per ottenere la risposta complessiva**; non si spegnerà mai un generatore continuo ed uno no, o entrambi accesi o entrambi spenti, perché altrimenti si altererebbe il punto di funzionamento statico.



Quindi, facendo riferimento alla figura, **il circuito a sinistra è l'amplificatore in sé e per sé**, che può essere **diviso in un circuito statico** (quello al centro) e in **un circuito dinamico** (quello a destra); **l'analisi statica**, che viene eseguita sul circuito statico, porta alla **determinazione del punto di funzionamento** precedentemente affrontata, mentre **per il circuito dinamico possono essere fatte le approssimazioni di piccolo segnale e semplificare il circuito in una sua versione lineare**, che non fa più utilizzo di transistori bipolaris a giunzione (non lineari) ma di generatori controllati:



Il circuito è descritto dalle stesse relazioni dinamiche dell'amplificatore precedente:

$$i_c(t) = g_m v_{be}(t)$$

$$i_b = \frac{v_{be}(t)}{r_\pi}$$

$$i_e = i_c + i_b = v_{be}(t) \left[\frac{1}{r_\pi} + g_m \right] = v_{be}(t) \left[\frac{g_m}{\beta} + g_m \right] = \frac{g_m}{\alpha} v_{be}(t) = \frac{v_{be}(t)}{r_e}$$

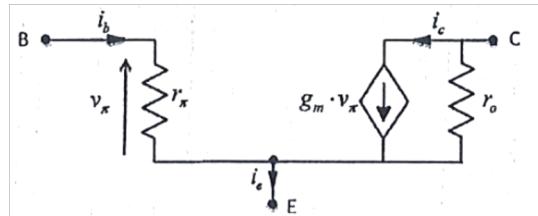
Si noti che, pur non comparendo, è presente anche la resistenza r_e . Quello appena descritto è il **modello equivalente per piccolo segnale del transistor bipolare a giunzione** e il circuito che si ottiene sostituendo al BJT tale modello è detto **circuito equivalente per piccolo segnale a π** .

Sebbene il circuito equivalente per piccolo segnale a π sia da utilizzare esclusivamente durante l'analisi dinamica (alias, "dal punto di vista del segnale"), esso **non è indipendente dal punto di funzionamento**; infatti, i **parametri caratteristici (g_m e r_π) sono un risultato della linearizzazione del punto di funzionamento e proporzionali alla corrente di polarizzazione**.

Finora, i **modelli equivalenti non considerano alcuna dipendenza della corrente di uscita dalla tensione di uscita perché durante la linearizzazione è stato trascurato l'effetto Early**; un modello

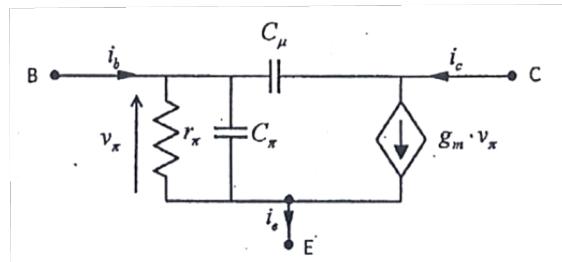
meno approssimato prevede che ci sia **un resistore in parallelo al generatore in ingresso di resistenza r_o** :

$$r_o = \left(\frac{di_C(t)}{dV_{CE}} \right)^{-1} = \frac{V_A}{I_C^*}$$



Di qui in avanti verrà sostituita la notazione $v_{be}(t)$ con $v_\pi(t)$ perché, a rigore, la tensione applicata ai terminali di base ed emettitore differisce dalla tensione che insiste sulla relativa giunzione a causa di effetti resistivi legati ai percorsi delle correnti; sebbene non vengano considerati questi effetti, è bene almeno usare la corretta nomenclatura.

Il modello semplificato appena mostrato viene usato quando al transistore sono applicati segnali variabili nel tempo. In determinate circostanze, soprattutto per segnali ad alte frequenze, nel transistore non possono essere trascurati alcuni effetti parassiti che si manifestano con un comportamento capacitivo (vengono schematizzati con condensatori ma non sono componenti fisici); tale comportamento si può osservare in seguito all'accumulo di cariche (sia fisse che mobili) in corrispondenza delle giunzioni base – emettitore e base – collettore quando aumenta la tensione (come effetto dell'aumento della frequenza e, quindi, dell'impedenza) e di conseguenza la regione di svuotamento (in corrispondenza della quale si accumulano ioni, quindi cariche). **Gli effetti capacitivi in questione possono essere modellati tramite due capacità** (variazione di carica nell'unità di tensione) poste tra i terminali corrispondenti:

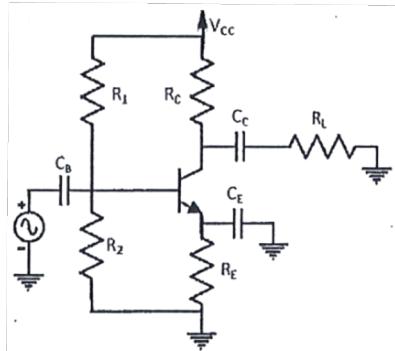


Le capacità C_μ e C_π sono molto piccole, nell'ordine dei pF , in modo che per basse frequenze possono essere modellate come circuiti aperti; la definizione di basse e alte frequenze sarà oggetto di una trattazione futura.

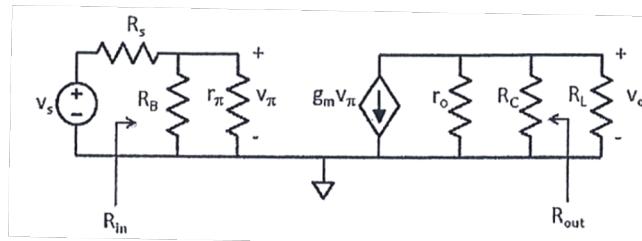
L'AMPLIFICATORE AD EMETTITORE COMUNE A BJT

Il dispositivo progettato finora prende il nome di **amplificatore ad emettitore comune** e, come il nome suggerisce, **presenta il nodo di emettitore in comune tra ingresso e uscita**; questo, viene collegato dinamicamente a massa dal condensatore C_E (supposto per il momento di capacità infinita affinché sia un cortocircuito ideale) in modo che il segnale di ingresso risulti collegato tra base ed emettitore ed il carico tra collettore ed emettitore.

Lo schema circuitale completo (e semplificato) di un amplificatore di questo tipo è stato già visto ma lo si ripropone nella figura seguente:



Nella trattazione che segue **si suppone di aver già determinato il punto di funzionamento statico del circuito** (quindi il valore delle quattro resistenze) e **che il segnale sia piccolo abbastanza da poter rientrare nelle condizioni di piccolo segnale**, affinché possa essere usato il circuito equivalente dinamico proposto di seguito:



La dinamica del circuito non prevede alcune grandezze tempo – invarianti, i generatori di tensione e i condensatori (quale che sia la frequenza del segnale, visto che si suppone la capacità infinita) sono stati sostituiti con cortocircuiti. **Gli effetti capacitivi del transistore sono ignorati finché la frequenza del segnale non risulta sufficientemente elevata**, condizione che per il momento si lascia da parte.

Fatte le dovereose premesse, **si passi ad analizzare il circuito dal punto di vista “del segnale”** (quindi **analisi dinamica**) e, in primis, **il guadagno dell’amplificatore**. Per **guadagno di tensione** si intende il **rapporto tra la tensione in uscita e la tensione in ingresso**; la prima è determinata dal fluire della corrente $g_m v_\pi(t)$ nel parallelo delle tre resistenze che si incontrano tra collettore e massa:

$$v_o(t) = -g_m v_\pi(t) (r_0 || R_C || R_L)$$

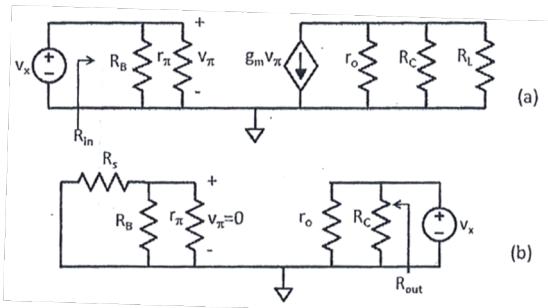
Tramite un semplice partitore resistivo nella maglia di ingresso, si può legare la tensione $v_\pi(t)$ alla tensione del generatore $v_s(t)$:

$$v_\pi(t) = v_s(t) \frac{r_\pi || R_B}{r_\pi || R_B + R_s}$$

Unendo i due risultati:

$$A_V = \frac{v_o(t)}{v_s(t)} = -\frac{g_m v_s(t) \frac{r_\pi || R_B}{r_\pi || R_B + R_s} (r_0 || R_C || R_L)}{v_s(t)} = -g_m \frac{r_\pi || R_B}{r_\pi || R_B + R_s} (r_0 || R_C || R_L)$$

Si completi l'analisi dell'amplificatore ad emettitore comune **valutando l'effetto che il circuito ha sulle resistenze di ingresso e di uscita**; per **resistenza di ingresso** si intende ciò che l'amplificatore mostra al generatore di segnale di ingresso, mentre per **resistenza di uscita** ciò che si vede dal carico guardando verso l'amplificatore. Il **calcolo della resistenza equivalente** tra due punti qualsiasi di un circuito viene effettuato **ponendo tra di essi un generatore test V_x , con la valutazione della corrente I_x da esso erogata** (avendo cura di spegnere gli altri generatori):



Nel calcolare R_{in} ci si riferisce al primo dei due circuiti test:

$$R_{in} = \frac{V_x}{I_x} = r_\pi || R_B$$

Per la resistenza R_{out} , essendo spento il generatore di segnale, la tensione $v_\pi(t) = 0$ e anche il generatore controllato viene spento. Per il calcolo ci si riferisce al secondo dei circuiti test in figura:

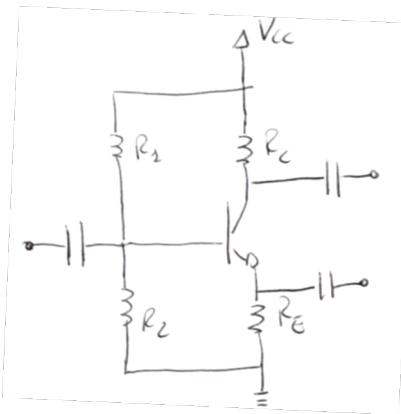
$$R_{out} = \frac{V_x}{I_x} = R_C || r_o$$

AMPLIFICATORI A COLLETTORE E A BASE COMUNE A BJT

Come si può intuire, l'**amplificatore ad emettitore comune non è l'unico amplificatore a BJT possibile**; in particolare, **in funzione di quale terminale del BJT è in comune tra ingresso e uscita** (e quindi dove sono posti segnale di ingresso e resistenza di carico), **si possono individuare tre tipologie di amplificatori**:

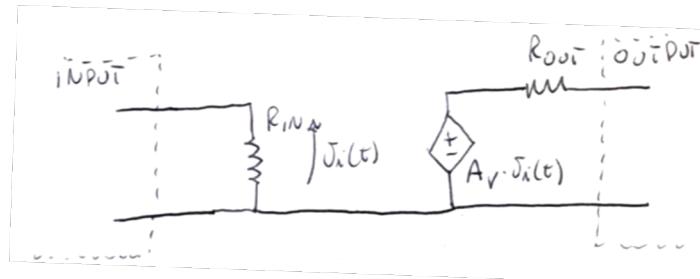
	Segnale di ingresso	Resistenza di carico
CE (Common Emitter)	Base	Collettore
CC (Common Collector)	Base	Emissore
CB (Common Base)	Emissore	Collettore

I tre hanno in comune il **circuito di polarizzazione**, dal quale si parte nella progettazione di un qualsiasi tipo di amplificatore a BJT; il **circuito non è unico**, quello precedentemente adoperato per introdurre la resilienza e limitare il numero di alimentazioni fa uso di quattro resistenze ed è configurato come segue:

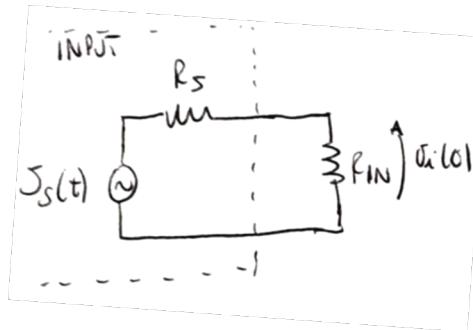


Vanno aggiunti dei condensatori di disaccoppiamento che distinguono il comportamento stazionario (DC) da quello dinamico (AC) e il cui valore condizionerà la risposta in frequenza.

Un amplificatore di tensione ideale può essere schematizzato con il seguente circuito equivalente, tenendo bene in considerazione che esso non ha alcun legame con il circuito equivalente a piccolo segnale (se non la forma); quest'ultimo serve unicamente a schematizzare il comportamento dinamico del BJT nell'ipotesi di piccolo segnale, stimando "carta e penna" i valori di R_{in} , R_{out} e A_V .



Il generatore controllato preleva la tensione $v_i(t)$ sulla resistenza R_{in} e la porta sulla maglia di uscita amplificata; sulla sinistra c'è l'input, su cui è applicato il segnale elettrico da amplificare e schematizzato con un generatore reale di tensione:



$$v_i(t) = v_s(t) \frac{R_{in}}{R_s + R_{in}}$$

Poiché il circuito amplifica $v_i(t)$, se si avesse $R_{in} = R_s$ si dimezzerebbe l'amplificazione complessiva.

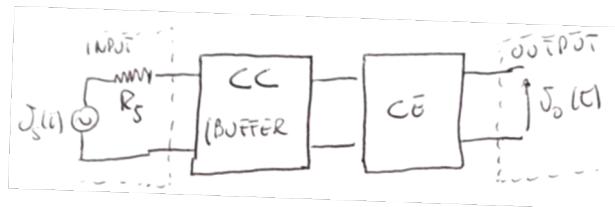
Per l'amplificatore ad emettitore comune (CE) sono stati individuati tre parametri: A_V , R_{in} e R_{out} . Per gli amplificatori CC e CB, gli stessi parametri sono riassunti nella seguente tabella (non si effettuano i calcoli per una questione di praticità):

	A_V	R_{in}	R_{out}
CC	< 1 (attenua)	Molto Grande	Molto piccola
CB	$g_m R_C$	Molto piccola	R_C

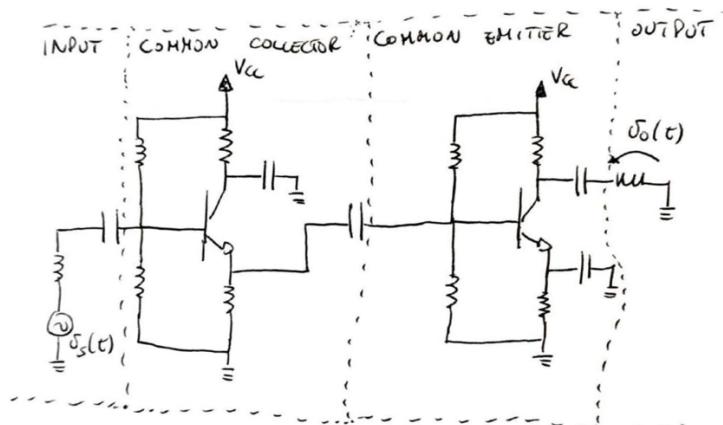
Configurazione	A_V	R_{in}	R_{out}
EC	$-g_m R_C$	r_π	R_C
BC	$g_m R_C \frac{R_{in}}{R_S + R_{in}}$	$\frac{1}{g_m}$	R_C
CC	< 1	$r_\pi + R_E (\beta_F + 1)$	$\frac{1}{g_m}$

Si noti che il guadagno del CB è positivo, a differenza del CE in cui figura il segno meno davanti; ciò rende l'amplificatore CB un amplificatore non invertente.

Ci si chiede che utilità abbia un amplificatore che non amplifica (cioè il CC, che ha guadagno prossimo ma minore all'unità); in realtà, l'amplificatore CC funge da buffer di tensione. Si supponga di avere a che fare con un segnale di ingresso caratterizzato da una R_s molto grande; se si volesse amplificare un segnale di questo tipo con un CE si perderebbe parte dell'amplificazione complessiva a causa del partitore di tensione fra R_{in} e R_s . La soluzione per ridurre le perdite sta nell'interposizione di un buffer di tensione tra il segnale di ingresso e l'amplificatore CE, in cui l'alta resistenza di ingresso R_s assicura che tutto il segnale cada su R_{in} e in cui la bassa resistenza di uscita R_{out} assicura che tutto il segnale cada sull'amplificatore CE.



Nella pratica, il CC serve come interfacciamento tra il segnale (prelevato dal mondo fisico) e un amplificatore CE; quello disegnato è uno schema black box che, concretizzato, si figura come segue:



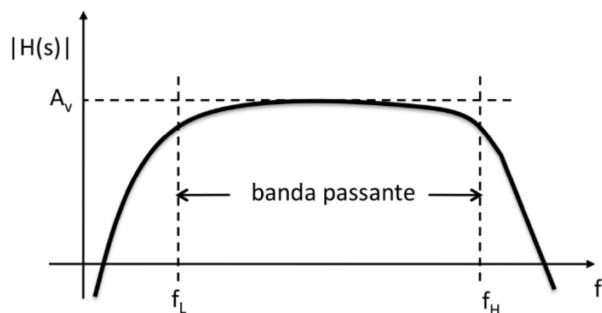
Ed è detto **doppio stadio CC – CE**.

Sulla base di quanto appena detto, **quando serve il CB?** La piccola resistenza di ingresso R_{in} rende questo amplificatore poco “fruibile” per l’amplificazione di tensione; infatti, è adoperato nel merito di amplificatori di corrente a BJT, che vanno fuori gli scopi della trattazione, interessa solo sapere che hanno una funzione duale al CC però relativa all’amplificazione di una corrente (quindi è un buffer di corrente).

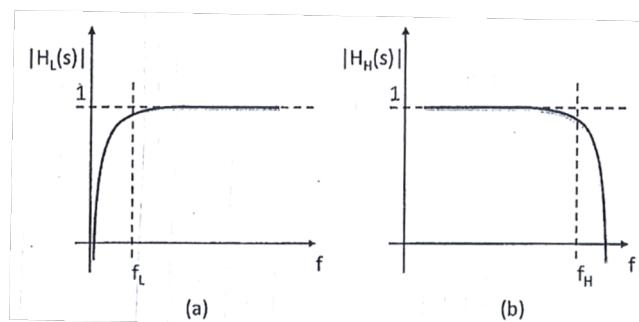
LA RISPOSTA IN FREQUENZA DEGLI AMPLIFICATORI A BJT

Nelle analisi effettuate sulle configurazioni elementari di amplificatori a BJT **non sono state tenute in conto le limitazioni in frequenza poste dagli elementi reattivi interni al BJT** o da quelle che si verificano quando i condensatori utilizzati per separare la polarizzazione dalla parte dinamica hanno valori finiti di capacità. In questa sede **non verrà fatta un’analisi rigorosa in un dominio trasformato** (Fourier o Laplace), sia per snellire la trattazione che **per fornire una stima non di quanto ma di dove il circuito è dipendente dalla frequenza**, individuando i componenti e le parti che sono responsabili delle limitazioni precedentemente introdotte.

In condizioni di piccolo segnale (la risposta in frequenza è una prerogativa dei sistemi lineari), si riporti su un grafico la dipendenza del guadagno di amplificazione in relazione alla frequenza del segnale:



Si osserva l’esistenza di una regione intermedia, detta **regione delle medie frequenze**, in cui il **guadagno è pressoché costante** (cioè **indipendente dalla frequenza del segnale**), e due regioni di **contorno**, individuate dalle **frequenze f_L e f_H** , dette rispettivamente **frequenza di taglio inferiore e superiore**. Per definizione, queste frequenze sono quelle oltre le quali il guadagno si riduce di **3dB** rispetto al valore alle medie frequenze, mentre l’intervallo di frequenze individuato tra le due ($f_H - f_L$) è detto **banda passante del circuito**. Poiché nella maggior parte dei circuiti elettronici $f_H \gg f_L$ (si dice che il circuito è a larga banda), il grafico mostrato può essere visto come la sovrapposizione dei due riportati di seguito:

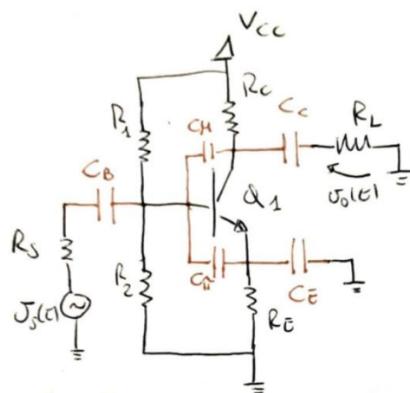


Di conseguenza, è possibile studiare la risposta in frequenza di un amplificatore elettronico separando il suo comportamento alle basse dal suo comportamento alle alte frequenze, visto che i fenomeni sono autoesclusivi; quindi, la funzione di trasferimento dell'amplificatore elettronico è:

$$H(f) = \frac{V_{out}(f)}{V_{in}(f)} = A_v H_L(f) H_H(f)$$

Dai grafici si può intuire che il comportamento alle basse frequenze è un comportamento passa – alto, mentre quello alle alte frequenze un passa – basso. Generalmente i condensatori di accoppiamento e bypass sono responsabili del comportamento passa – alto, mentre le capacità interne del BJT sono responsabili del comportamento passa – basso.

Considerati anche gli effetti capacitivi del transistore (si ricordi che non sono capacitorsi in sé e per sé, la schematizzazione li porta per rappresentare effetti capacitivi), l'amplificatore a BJT assume la seguente forma:

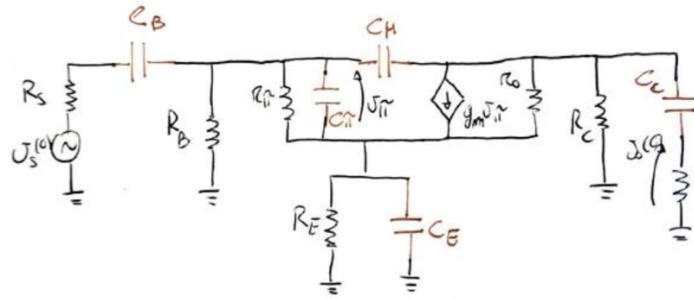


Si è di fronte a cinque capacità: C_B , C_C , C_E , C_π e C_μ ; le prime tre corrispondono a dei componenti piazzati ad hoc per poter avere l'accesso diretto ai terminali di base, collettore ed emettitore dal punto di vista dinamico, le rimanenti due sono capacità parassite (in RAD, C_π capacità di diffusione, relativa ad una giunzione polarizzata, e C_μ capacità di giunzione, relativa ad una giunzione polarizzata inversamente), non possono essere dimensionate al di fuori del processo produttivo del BJT (in particolare, sono associate alle giunzioni PN che si trovano nel componente, C_π alla giunzione base – emettitore e C_μ alla giunzione base – collettore). Dal punto di vista dinamico, si ha margine solo sulle prime tre, dato che sono componenti reali, con le altre bisogna convivere.

Di seguito è schematizzato il comportamento delle due tipologie di capacità in un amplificatore a BJT:

	C_B , C_C e C_E	C_π e C_μ
Dimensione	Nell'ordine dei μF a salire	Nell'ordine dei pF
Effetto in frequenza	Definiscono la frequenza di taglio inferiore ω_L	Definiscono la frequenza di taglio superiore ω_H

Per vedere cosa succede al circuito sopra mostrato bisogna porsi in condizione di piccolo segnale e considerare frequenze $f = \omega/2\pi > 0$; analizzando il comportamento dinamico (per cui tutti i generatori continui sono cortocircuitati), si ottiene la seguente configurazione:



Che vale per ogni frequenza. Chiaramente, il circuito è molto complesso e lo studio si può effettuare solo ponendosi lontano le frequenze di taglio; quindi, **si individuano le tre regioni possibili e si distinguono i comportamenti delle varie capacità:**

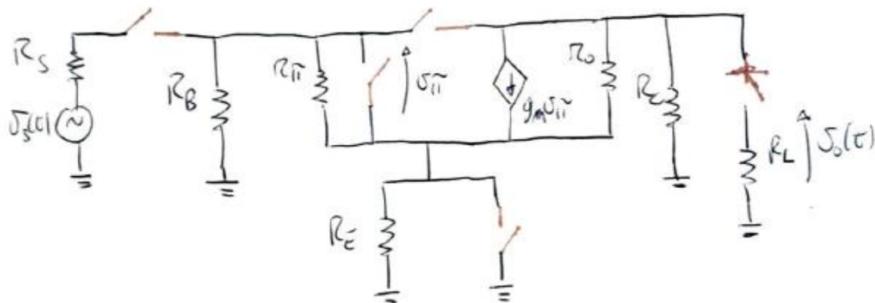
	$\omega \ll \omega_L$	$\omega_L \ll \omega \ll \omega_H$	$\omega > \omega_H$
C_B, C_C e C_E	APERTI	CORTI	CORTI
C_π e C_μ	APERTI	APERTI	CORTI

Ovviamente la notazione è puramente didattica, visto che sono i valori delle capacità a determinare ω_H e ω_L ; in questa sede non verranno approfonditi, però, i calcoli e le formule per determinare tali frequenze per una questione di semplicità.

Si studi il circuito per ognuna delle tre regioni individuate:

- $\omega \ll \omega_L$

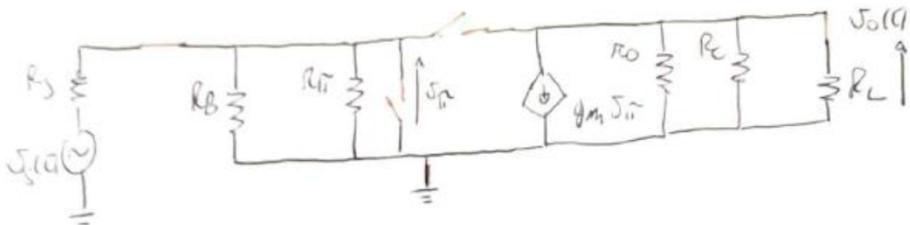
Alle basse frequenze il circuito assume la forma:



Si noti che il generatore di segnale non è nemmeno fisicamente collegato al circuito amplificatore e non ci si può aspettare che riesca a condizionare $v_\pi(t)$ in modo da far scorrere corrente nel generatore controllato ed amplificare; anche qualora accadesse, l'uscita è fisicamente scollegata dal circuito, l'aperto impone una corrente nella resistenza di carico che è necessariamente nulla e, quindi, $v_o(t) = 0$. L'amplificatore non sta funzionando.

- $\omega_L \ll \omega \ll \omega_H$

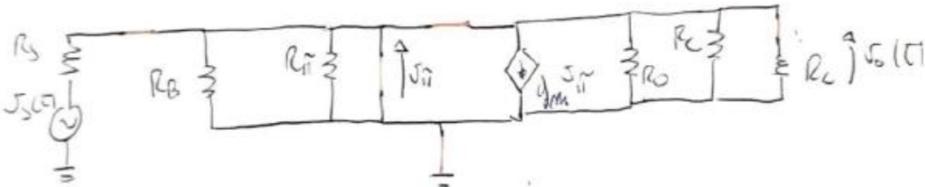
Alle medie frequenze il circuito assume la forma:



Corrisponde al circuito equivalente analizzato nei precedenti capitoli; l'emettitore è cortocircuitato a massa, mentre l'ingresso e l'uscita sono connessi al circuito. Tutto funziona correttamente e il circuito amplifica; quindi, si può desumere che lo studio degli amplificatori fatto fino a questo momento era da intendere come studio alle medie frequenze.

- $\omega \gg \omega_H$

Alle alte frequenze il circuito assume la forma:



Tutte le capacità sono diventate cortocircuiti, in parallelo a r_π si trova un corto e ai suoi capi si trova $v_\pi(t)$; tuttavia, la corrente che scorre nel generatore controllato è nulla in quanto il corto su r_π impone una tensione $v_\pi(t) = 0$ e, inoltre, il corto su C_π (in concomitanza a quello su C_μ) cortocircuita anche la resistenza di carico R_L , rendendo l'uscita $v_o(t) = 0$. Similmente alle basse frequenze, l'amplificatore non sta funzionando.

Adesso ci si chiede cosa succede alle medie frequenze; si consideri un segnale a bassa frequenza, incrementandola pian piano. Per prima cosa, i capacitori C_B , C_C e C_E iniziano a ridurre la propria impedenza ($|Z_C| = 1/\omega C$) fino a diventare corti; la stessa cosa non si può dire per C_π e C_μ , dal momento in cui il loro valore di capacità è talmente basso (nell'ordine dei pF) che lo stesso effetto è raggiunto molte decadi dopo in termini di frequenza (cioè a frequenze centomila/un milione di volte più alte). In queste condizioni ci si ritrova nella regione in cui il circuito amplifica, andando avanti fino a che la frequenza non diventa sufficientemente alta da ridurre anche le impedenze di C_π e C_μ ; a tal punto, gradualmente il guadagno inizia a diminuire fino ad annullarsi.

In questo modo è giustificato il comportamento passa – banda del circuito amplificatore rilevato ad inizio trattazione. Essendo il guadagno costante (e si potrebbe anche notare la fase proporzionale alle frequenze) nella banda, il circuito amplificatore non effettua distorsione; inoltre, nella maggior parte delle misurazioni si nota un guadagno negativo ed il perché risiede nel fatto che quello è il modulo di un numero complesso la cui fase è in un intorno di $-\pi$.

A_v è detto guadagno di amplificazione alle medie frequenze proprio perché è il valore rilevato quando le frequenze sono abbastanza alte da annichilire le impedenze di C_B , C_C e C_E ma non abbastanza da annichilire quelle di C_π e C_μ . Qualitativamente, le frequenze di taglio sono ottenute riducendo il guadagno di 3dB e prendendo le ascisse dei punti corrispondenti a quel valore; nel caso in cui si volesse aumentare o ridurre la banda passante, si dovrebbe agire su:

- C_B , C_C e C_E in aumento per diminuire ω_L ;
- C_π e C_μ in diminuzione per aumentare ω_H .

Tuttavia, per il primo intervento ci si deve chiedere se valga la pena, visto che capacitori con capacità più elevate costano e sono ingombranti, mentre per il secondo intervento non si ha margine se non nella scelta di transistori appositamente progettati (su C_π e C_μ non si ha diretto margine). In ogni caso, la modifica della banda passante ha senso se e solo se i segnali che si trattano hanno frequenze che non appartengono alla attuale banda.

Uno dei transistori che forniscono valori di C_π e C_μ più piccoli è l'HJT (Heteropolar Junction Transistor) basato su tecnologie che sfruttano materiali composti (come silicio e germanio).

Si vuole concludere menzionando le **due possibili rappresentazioni del guadagno**, quella in tensione su tensione e in decibel. La prima viene calcolata semplicemente considerando il rapporto tra tensione in uscita e tensione in ingresso all'amplificatore ed è utile per una stima umanamente più comprensibile dell'amplificazione, mentre la seconda è più utile per determinare le **frequenze di taglio** e si calcola:

$$A_{dB} = 10 \cdot \log_{10} A_v$$

Nel caso in cui la conversione fosse fatta sulla base delle potenze (come in LTSpice) e ricordando che la formula della potenza ha la tensione elevata al quadrato:

$$A_{dB} = 10 \cdot \log_{10} A_v = 10 \cdot \log_{10} \frac{v_{out}(t)^2}{v_{in}(t)^2} = 10 \cdot \log_{10} \left(\frac{v_{out}(t)}{v_{in}(t)} \right)^2 = 20 \cdot \log_{10} A_v$$

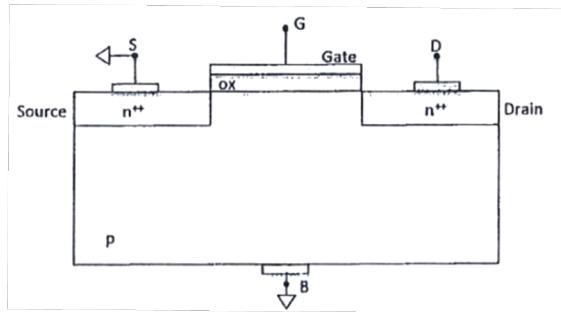
Il decibel è una rappresentazione anche familiare al modo in cui funziona l'orecchio umano: quando un'onda meccanica giunge all'orecchio ad una ampiezza doppia rispetto al normale, non è percepita il doppio più intensa, mentre lo sarebbe se fosse ad una ampiezza decuplicata.

IL MOSFET

Sebbene l'era dei dispositivi elettronici sia convenzionalmente iniziata con l'invenzione del BJT nel 1947, circa vent'anni prima (negli anni 20') J.E. Lilienfeld aveva scoperto e brevettato un effetto fisico, detto effetto di campo, per il quale era possibile regolare la corrente circolante tra due terminali di un dispositivo a stato solido mediante un'ulteriore terminale di controllo; tuttavia, a causa di limiti tecnologici, i FET (Field Effect Transistor) furono sviluppati e ingegnerizzati solo tra gli anni 70' e 80', sorpassando gradualmente l'impiego dei BJT.

Lo stesso BJT fu sviluppato con l'intenzione di creare un FET ma Shockley, Brattain e Bardeen commisero un errore che gli valse il premio Nobel e che permise la prosecuzione delle tecnologie necessarie per creare ciò che loro stessi volevano ingegnerizzare. La differenza tra i due dispositivi è nel luogo in cui si sviluppa il funzionamento: nei FET avviene sulla base della qualità della superficie del semiconduttore (infatti sono detti **dispositivi superficiali**), mentre nei BJT le caratteristiche sono dovute a ciò che accade nella base, che è all'interno del semiconduttore. Differenze a parte, ad oggi nella versione attuale (MOSFET, Metal Oxide Semiconductor Field Effect Transistor) il dispositivo di Lilienfeld ricopre più del 90% di tutte le applicazioni in cui è necessario l'effetto transistore.

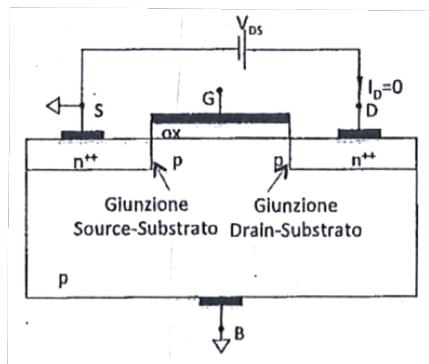
Lo schema in sezione di un dispositivo MOSFET è descritto dalla seguente figura:



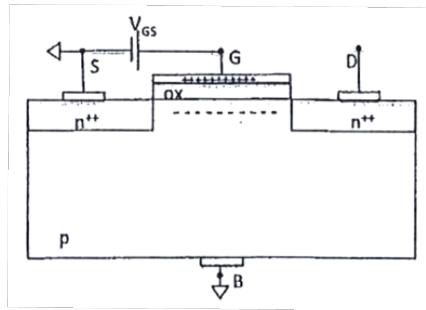
In un substrato di tipo P vengono realizzate due regioni superficiali di tipo N molto drogate, la regione di **Source** (S) e di **Drain** (D), tra le quali è presente uno strato superficiale di Ossido di Silicio SiO_2 (o Nitruro di Silicio Si_3N_4) a sua volta coperto da uno strato metallico, il **Gate** (G). Verticalmente, si può osservare una struttura simile ad un condensatore a facce piane e parallele, costituito da **Metallo – Ossido – Semiconduttore** in cui l'ossido funge da dielettrico e metallo e semiconduttore da armature. Generalmente, il **MOSFET** è caratterizzato da quattro terminali, **S – D – G** e un terminale che contatta il substrato, il **Body** (B), che verrà lasciato a terra come potenziale di riferimento.

Nell'analisi che verrà fatta a breve, si suppongono i terminali di **Source** e **Body** allo stesso potenziale (di riferimento) mostrando come è possibile gestire il passaggio di corrente tra **Drain** e **Source** tramite l'applicazione di una tensione al **Gate** (ponendo il **Body** a terra, lo si impone al potenziale minore e si polarizza inversamente la giunzione); quindi, in maniera del tutto analoga al **BJT**, verrà mostrato come la corrente in un terminale dipende unicamente dalla tensione applicata in una regione del dispositivo separata, restituendo un comportamento da generatore controllato di corrente.

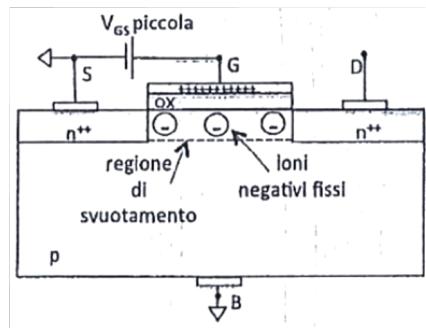
Si supponga inizialmente di non applicare alcuna tensione al **Gate** e si verifichi se, applicando una $V_{DS} > 0$ tra **Drain** e **Source**, sia possibile avere una circolazione di corrente tra questi due terminali:



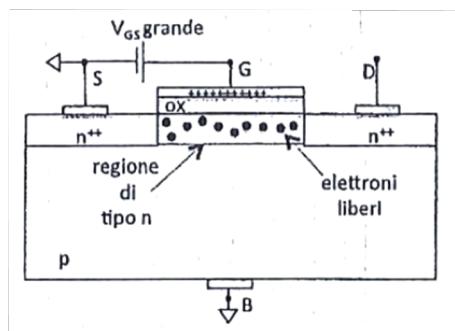
Affinché la corrente I_D circoli, è necessario che attraversi le giunzione **Source – Substrato** e **Drain – Substrato**; indipendentemente dal verso della tensione applicata, una delle due giunzioni sarà polarizzata inversamente, impedendo qualsiasi circolazione di corrente e restituendo $I_D = 0$. Invece, lasciando non alimentato il terminale di **Drain**, si applichi una tensione $V_{GS} = V_G > 0$; ci si aspetta un comportamento capacitivo, quindi un accumulo di carica sul terminale di **Gate** e sul substrato in accordo con la tensione applicata, secondo la relazione $Q = CV$:



Nel caso considerato, il substrato è di tipo P, quindi ricco di lacune, e l'accumulo di carica negativa in superficie avviene con due modalità diverse, che dipendono dall'intensità della tensione applicata. In un primo momento, per tensioni basse, il campo elettrico che si instaura tra le due armature allontana le lacune prossime all'interfaccia con l'ossido e le cariche negative accumulate sono gli atomi droganti, non più neutri perché l'allontanamento delle lacune li ha resi ioni negativi; il fatto che le cariche accumulate siano ioni (quindi fissi) fa sì che in superficie si generi una regione di svuotamento:



Quando aumenta V_{GS} , vengono richiamati verso la superficie gli elettroni liberi che si trovano nelle regioni di tipo P circostanti; tali elettroni vanno ad invadere la regione al di sotto dell'ossido, fornendo energia sufficiente al bilanciamento della tensione applicata. Essendo gli elettroni cariche elettriche mobili, la regione superficiale del semiconduttore non è più di tipo P ma è diventata di tipo N; si dice che è avvenuta la cosiddetta inversione di popolazione e le regioni di Drain e Source sono collegate tra di loro attraverso un'ulteriore regione, detta regione di canale, ancora di tipo N:



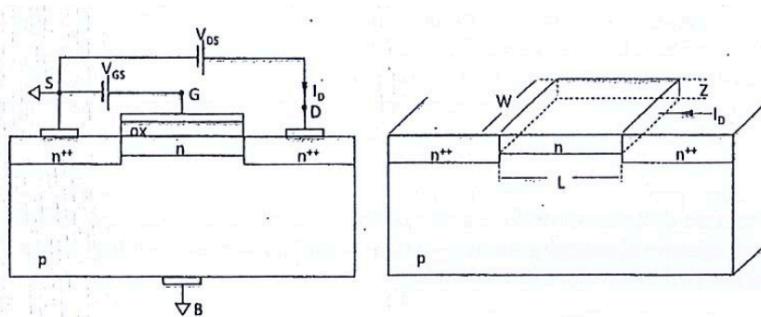
L'esistenza di una regione dello stesso tipo di Source e Drain permetterebbe ad un'eventuale corrente circolante tra i due terminali di non essere bloccata da alcuna giunzione PN polarizzata inversamente, dal momento in cui vedrebbe solo un unico percorso omogeneo di tipo N.

LE CARATTERISTICHE TENSIONE – CORRENTE DI UN MOSFET

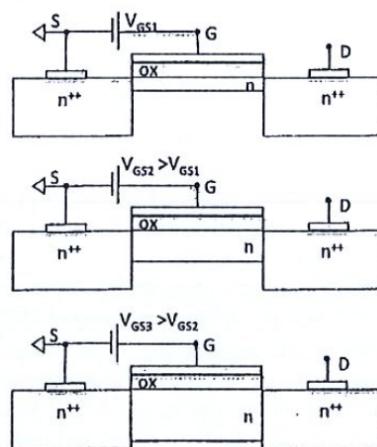
In presenza di una $V_{GS} > 0$ sufficientemente elevata, si applichi una $V_{DS} > 0$ in modo da far scorrere una corrente $I_D \neq 0$. La tensione necessaria a rilevare tale corrente è detta **tensione di soglia** V_{th} (th sta per threshold) ed è definita quantitativamente come la tensione per la quale la **concentrazione della carica libera invertita** (elettroni in questo caso) uguaglia la **concentrazione delle cariche del substrato** (lacune in questo caso). Per $V_{GS} > V_{th}$ la concentrazione di elettroni richiamati nella regione di canale aumenta, portando ad un **conseguente aumento della conducibilità**; la resistenza offerta dal canale al passaggio della corrente I_D dipende dalla **conducibilità e dalle geometrie della struttura**:

$$R = \frac{1}{\sigma} \frac{L}{WZ}$$

Con **L** la lunghezza del canale, **W** la dimensione trasversale e **Z** l'ampiezza, ossia la distanza dall'interfaccia ossido – silicio entro cui sono confinati gli elettroni:

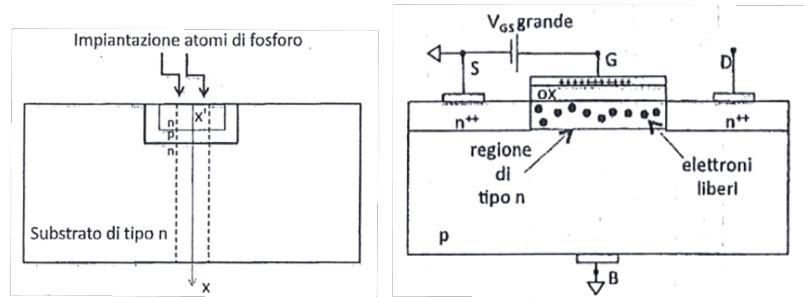


Nella realtà, **Z** è una grandezza fittizia, la regione in cui gli elettroni si addensano è così piccola che può essere considerata sostanzialmente nulla; tuttavia, per una semplicità di notazione, si preferisce considerare la **conducibilità σ** invariante rispetto a V_{GS} e che la diminuzione della resistenza complessiva (attribuita al maggior numero di cariche richiamate verso l'interfaccia) sia da associare ad una maggiore estensione **Z** del canale:

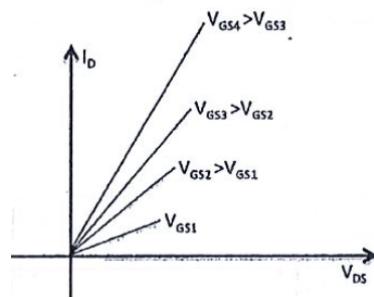


Nel calcolo della resistenza complessiva tra i terminali di Drain e di Source non viene tenuto in conto il contributo delle regioni di tipo N a cui i terminali sono associati ed il motivo è che, essendo molto drogate (N^{++}), si osserverebbe una conducibilità prossima a quella di un metallo ed una resistenza pressoché nulla, quindi trascurabile.

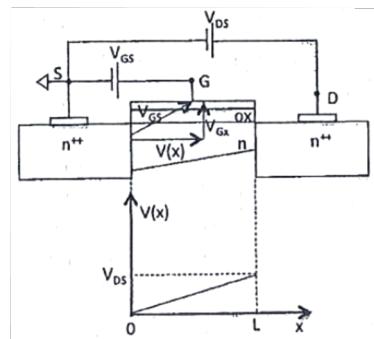
Prima di procedere verso ragionamenti analitici, ci si vuole soffermare su una delle principali differenze tra i BJT e i MOSFET: nei primi la corrente scorre verticalmente, tra collettore ad emettitore attraverso il terminale di base, nei secondi scorre orizzontalmente, tra Drain e Source attraverso la regione di canale.



Fatte tutte queste premesse, è ora di procedere a descrivere l'andamento della corrente di Drain I_D in funzione della tensione applicata tra i terminali di Drain e Source V_{DS} e prendendo come parametro la tensione applicata al terminale di Gate V_{GS} . Poiché la tensione V_{DS} è applicata direttamente al canale e poiché questo ha un comportamento resistivo, è lecito aspettarsi una corrente I_D che varia linearmente; a parità di V_{DS} , la caratteristica è più pendente quando il parametro V_{GS} è maggiore, dal momento in cui regola l'ampiezza del canale e la diminuzione della resistenza. In una prima fase le caratteristiche assumono la seguente forma:



Quando si vanno a considerare tensioni V_{DS} non più tanto piccole, il comportamento lineare non è più sufficiente a descrivere la relazione con la I_D ; per aggiornare le caratteristiche sono necessarie delle digressioni sulla maniera in cui la tensione è distribuita all'interno del canale. Si consideri la situazione mostrata nella figura seguente, in cui V_{DS} è ancora bassa, il comportamento del canale è ancora resistivo e la tensione si distribuisce lungo di esso con una legge lineare:

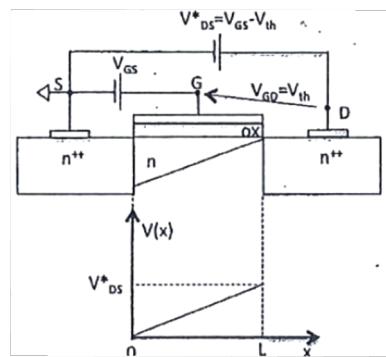


In un riferimento con l'origine in corrispondenza dell'interfaccia tra Source e canale, la figura mostra l'andamento della caduta di tensione lungo il canale tra 0 e L , quindi $V(x)$; in 61

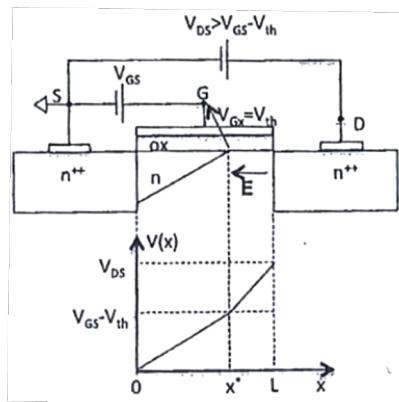
corrispondenza del Source, si ha $V(x) = 0$ mentre alla fine del canale $V(x) = V_{DS}$. È chiaro, quindi, che la differenza di potenziale V_{Gx} che insiste ai capi dell'ossido, determinata come differenza tra il potenziale al Gate ed il potenziale al punto x lungo il canale, **dipende dal punto x stesso in cui è considerata**, variando dal Source verso il Drain; ad esempio, per $x = 0$ si ha $V_{Gx} = V_{GS}$ e per $x = L$ si ha $V_{Gx} = V_{GD} = V_{GS} - V_{DS}$. Generalmente, la ddp ai capi del condensatore MOS in corrispondenza della sezione x è pari a:

$$V_{Gx} = V_{GS} - V(x)$$

La conseguenza è che l'ampiezza del canale non è costante andando dal Source verso il Drain: è massima in corrispondenza del Source e minima in corrispondenza del Drain; tuttavia, affinché si formi il canale conduttivo la ddp ai capi del condensatore MOS deve essere maggiore della tensione di soglia V_{th} . Questa condizione è imposta dall'esterno sulla V_{GS} ma non è garantito che sia verificata per ogni sezione lungo il canale; generalmente, per ogni V_{GS} esiste una V_{DS}^* al di sopra della quale $V_{GS} - V_{DS} < V_{th}$ (evidentemente $V_{DS}^* = V_{GS} - V_{th}$). In corrispondenza di questa tensione, detta tensione di saturazione (o di pinch-off), il canale conduttivo si strozza ed assume la seguente forma:



Quando la V_{DS} supera la tensione di pinch-off la corrente di Drain I_D rimane costante e per comprenderne il motivo bisogna fare riferimento alla seguente figura:



Quando $V_{DS} > V_{DS}^*$ esiste sicuramente un punto x^* tra 0 e L per cui $V(x^*) = V_{DS}^* = V_{GS} - V_{th}$; in questa configurazione, la tensione ai capi del condensatore risulta essere $V_{Gx} = V_{GS} - V(x) = V_{th}$ ed il canale si strozza in prossimità di x^* , delimitando due regioni: una da 0 a x^* , in cui sono presenti elettroni liberi, ed una da x^* a L , in cui non ci sono cariche libere. La dimensione della prima regione dipende da V_{DS} , la quale aumentando trascina x^* verso il Source.

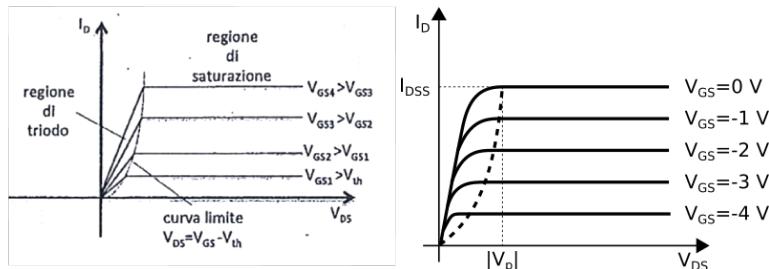
Ciò che va sottolineato è che **ai capi della prima regione la tensione sarà sempre $V_{GS} - V_{th}$** (detta **tensione di overdrive**), **dal momento in cui $V(x^*) = V_{DS}^*$ non dipende da essa**. Con una buona approssimazione, **è possibile dire che la resistenza associata a questa regione non varia al variare di V_{DS}** (nella realtà il contributo dato dalla regione compresa tra x^* e L è trascurabile, visto che tale distanza sarà quasi nulla):

$$R = \frac{V_{GS} - V_{th}}{I_D} \Rightarrow I_D = \frac{V_{GS} - V_{th}}{R}$$

Segue che **neanche la corrente dipende** (in maniera rilevante, una minima linearità c'è sempre ma è trascurabile) **dalla tensione V_{DS} ma, bensì, dalla tensione applicata a due terminali verso cui non è rivolta, V_{GS} .**

Resta da determinare cosa accade nella **regione trascurata**, cioè quella che va da x^* a L ; da 0 a x^* la tensione cade linearmente fino a raggiungere il valore $V_{GS} - V_{th}$ mentre nel tratto successivo viene scaricata la restante tensione, come la figura mostra adeguatamente. L'eccesso di tensione in questione insiste su una **distanza molto piccola**, dando luogo ad un **campo elettrico molto forte diretto dal Drain al Source**; la funzione di questo campo elettrico è **del tutto analoga a quella del campo presente sulla giunzione base – collettore di un BJT nei confronti degli elettroni provenienti dall'emettitore**: nel caso del MOSFET, gli elettroni partono dal **Source sotto l'influenza della tensione V_{DS}^* e giungono fino a x^* , dove il campo elettrico li raccoglie e li spinge ad attraversare la regione di svuotamento indirizzandoli verso il Drain**.

Il risultato che fuoriesce da queste considerazioni porta a determinare come segue la famiglia di caratteristiche di uscita di un MOSFET:



Si individuano le seguenti **regioni di funzionamento di un MOSFET**, sulla base delle caratteristiche in questione:

- **Regione lineare (o di triodo)**, presente per ogni V_{GS} fintantoché $V_{DS} < V_{GS} - V_{th}$:
 - Il legame tra I_D e V_{DS} è lineare, per poi addolcirsì in prossimità di $V_{DS} = V_{DS}^*$ con un andamento quadratico a causa dell'aumento della resistenza dovuta al restringimento del canale;
- **Regione di saturazione (o di pinch – off)**, presente per ogni V_{GS} fintantoché $V_{DS} > V_{GS} - V_{th}$:
 - Il legame tra I_D e V_{DS} è costante (in realtà è lineare ma con una pendenza minima);
 - Non va confusa con la regione di saturazione di un BJT, prossima all'asse delle correnti e più vicina alla regione di triodo;
- **Regione di interdizione**, coincidente con l'asse x e caratterizzata da una corrente nulla
 - È utile ricordare che $I_D = 0$ per ogni tensione $V_{GS} < V_{th}$ indipendentemente da V_{DS} .

Il dispositivo appena descritto è detto **MOSFET a canale N (o N – MOS) ad arricchimento**, indicando che la **regione di canale viene arricchita con cariche elettriche mediante l'applicazione della tensione V_{GS}** ; esiste il dispositivo speculare ma non verrà approfondito adesso.

Con queste informazioni, si possono stimare qualitativamente le caratteristiche di uscita di un **MOSFET a canale N**; per V_{DS} piccole, la corrente I_D è tanto proporzionale a V_{DS} quanto allo scarto di V_{GS} rispetto a V_{th} :

$$I_D \propto V_{DS}(V_{GS} - V_{th})$$

Quando V_{DS} si avvicina a V_{DS}^* per ogni singola V_{GS} , le caratteristiche deviano verso un andamento di tipo quadratico:

$$I_D = k[2(V_{GS} - V_{th})V_{DS} - V_{DS}^2]$$

Si nota che per V_{DS} piccole il termine V_{DS}^2 è trascurabile e la caratteristica è lineare; inoltre, il modello non descrive bene il MOSFET per $V_{DS} > V_{DS}^*$ perché la parabola ha il vertice in $V_{DS} = V_{GS} - V_{th}$ e per valori maggiori decresce, quando è stato rilevato che rimane pressoché costante. Tuttavia, l'equazione può servire per determinare il modello a $V_{DS} > V_{DS}^*$ perché può fornire il valore limite:

$$I_D(V_{DS} = V_{DS}^* = V_{GS} - V_{th}) = k[2(V_{GS} - V_{th})(V_{GS} - V_{th}) - (V_{GS} - V_{th})^2] = k(V_{GS} - V_{th})^2$$

Infatti, dal punto di raccordo in poi la corrente di Drain dipende solo dalla tensione $V_{GS} - V_{th}$ e non più da V_{DS} . In definitiva, le caratteristiche di uscita di un N – MOS sono:

$$\begin{cases} V_{GS} < V_{th} \Leftrightarrow I_D = 0 \\ V_{GS} > V_{th} \Leftrightarrow \begin{cases} I_D = k[2(V_{GS} - V_{th})V_{DS} - V_{DS}^2] \Leftrightarrow V_{DS} < V_{GS} - V_{th} \\ I_D = k(V_{GS} - V_{th})^2 \Leftrightarrow V_{DS} \geq V_{GS} - V_{th} \end{cases} \end{cases}$$

Per il BJT è stato necessario rilevare anche le caratteristiche di uscita per la corrente di Emettitore e di Base; per il N – MOS non è necessario, visto che la presenza dell'ossido di silicio impedisce il passaggio di una corrente di Gate e che, per necessità, la corrente di Source deve essere uguale (in modulo) a quella di Drain:

$$\begin{cases} I_G = 0 \\ |I_S| = |I_D| \end{cases}$$

Con il modello a disposizione, è possibile anche individuare l'espressione analitica della curva limite tra la **regione di triodo** e la **regione di pinch – off**, definita dall'uguaglianza $V_{GS} = V_{DS} + V_{th}$; sostituendo nell'equazione:

$$I_D = kV_{DS}^2$$

Che, nel piano $I_D - V_{DS}$ rappresenta la parabola mostrata nella figura precedente.

Nelle equazioni è apparsa una **costante di proporzionalità**, k , della quale non è stato ancora detto nulla; essa è una costante che prende le dimensioni di una conduttanza, può essere vista (ma non lo è in realtà) come l'inversa della resistenza che si trova nel canale tra Source e Drain e **dipende sia dalle geometrie del MOSFET che da parametri fisici legati alla regione di canale**:

$$k = \frac{1}{2} C_{ox} \mu_n \frac{W}{L}$$

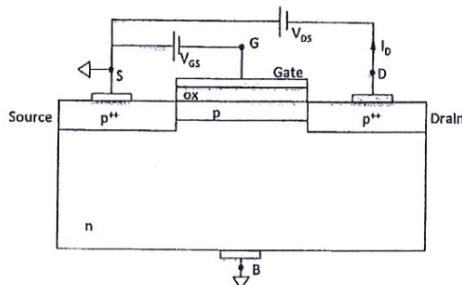
Dove W e L sono gli stessi parametri introdotti ad inizio trattazione, μ_n è la mobilità degli elettronni (perché si parla di N – MOS) e C_{ox} è la capacità del condensatore MOS, legata allo spessore e alla composizione dell'ossido di silicio:

$$C_{ox} = A \frac{\epsilon_{ox}}{t_{ox}}$$

A è l'unità di area nella regione di Gate e permette a k di variare in funzione della geometria del dispositivo (se fosse WL non accadrebbe), ϵ_{ox} la costante dielettrica dell'ossido e t_{ox} il suo spessore.

IL MOSFET A CANALE P E A SVUOTAMENTO

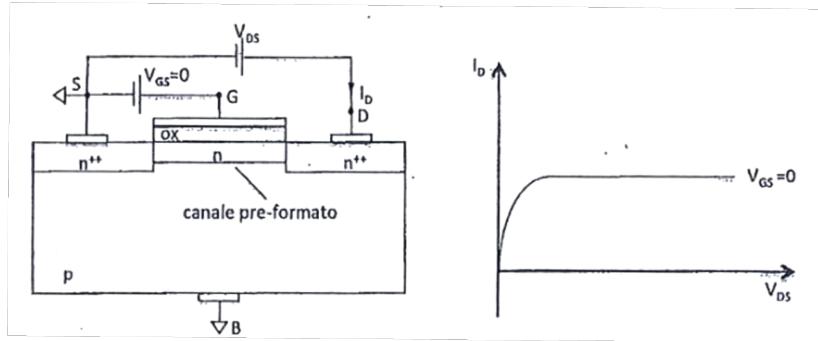
Come per il BJT NPN ha la sua versione duale PNP, il **MOSFET a canale N ha la sua versione duale nel MOSFET a canale P**, realizzato a partire da un substrato di tipo N poco drogato nel quale vengono realizzate due regioni superficiali di tipo P molto drogate. L'applicazione di una tensione V_{GS} negativa al terminale di Gate richiama cariche positive verso l'interfaccia e, per valori in modulo sufficientemente grandi, provoca l'inversione della popolazione: la concentrazione di lacune richiamate supera la concentrazione di elettroni nel substrato e genera un canale conduttivo di tipo P.



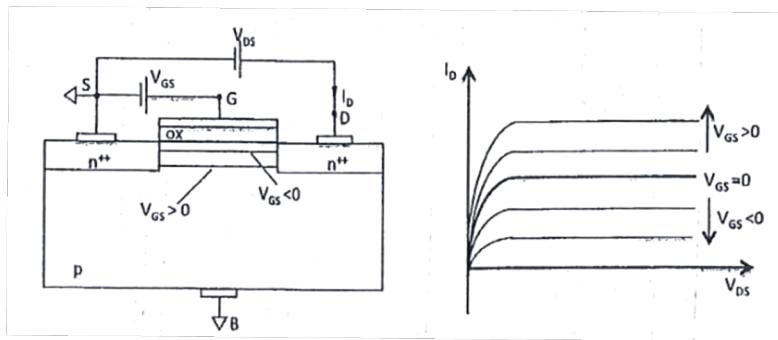
In linea di principio, il funzionamento di questo dispositivo è del tutto analogo a quello del suo duale, sebbene le equazioni analitiche che ne descrivono il comportamento saranno fornite solo in un successivo momento. Ciò che bisogna sapere, tuttavia, è che si verifica un fenomeno analogo a quello osservato con il BJT: per il MOSFET a canale P la dipendenza dalla mobilità della corrente di Drain impone un valore minore (in modulo assoluto) al proprio dispositivo duale.

Le immagini non sono propriamente accurate perché il Body non va a massa ma viene collegato all'alimentazione; infatti, se il Body fosse a massa il Source potrebbe essere a potenziale maggiore e ci potrebbe essere scorrimento di corrente, mentre la giunzione è polarizzata inversamente se il Body è direttamente a potenziale maggiore.

Prendendo come riferimento un MOSFET a canale N, si nota che la regione di canale non si forma per valori di $V_{GS} < V_{th}$, che a sua volta è positiva; quindi, per $V_{GS} = 0$ il canale non è formato. Esistono dispositivi, detti **MOSFET a svuotamento**, la cui regione di canale non viene realizzata con l'applicazione di una V_{GS} all'atto dell'utilizzo ma in fase di fabbricazione del dispositivo; ne segue che, per $V_{GS} = 0$, il canale è preformato.



Quindi, l'applicazione di una tensione V_{DS} provoca lo scorrimento di una corrente I_D anche quando $V_{GS} = 0$. Per valori di $V_{GS} > 0$ non si fa altro che richiamare più elettroni nella regione di canale, aumentando il valore di I_D proprio come in un MOSFET ad arricchimento; infatti, i due dispositivi sono uguali, l'unica differenza è il valore di tensione per cui il canale si forma, positivo per l'arricchimento e negativo per lo svuotamento. Per avvalorare quanto appena detto, si può osservare che le caratteristiche tensione – corrente di un MOSFET a svuotamento assumono la stessa forma delle caratteristiche di un MOSFET ad arricchimento, solo che il valore V_{GS} di una stessa curva è più piccolo nel primo dispositivo che nel secondo:



Procedendo a ritroso, per valori di V_{GS} negativa, il canale preformato andrà a restringersi sempre di più proprio come accadrebbe ad un MOSFET ad arricchimento se si procedesse da $V_{GS} > V_{th}$ verso valori prossimi a V_{th} ; la tensione per cui il canale cessa di esistere è detta tensione di svuotamento, V_{TD} , ed è quella da osservare quando si vuole verificare l'interdizione o la conduzione del dispositivo. È ovvio che V_{TD} e V_{th} rappresentano lo stesso stato del dispositivo, ovvero quello per cui inizia a formarsi il canale di conduzione, solo che in un MOSFET ad arricchimento si raggiunge V_{th} da valori minori (e quindi il canale non esistente si crea) e in un MOSFET a svuotamento si raggiunge V_{TD} da valori maggiori (e quindi il canale esistente si annichilisce).

Analiticamente, le caratteristiche di un MOSFET a svuotamento sono uguali a quelle di un MOSFET ad arricchimento, solo che andrà considerata la traslazione delle curve rispetto alla V_{GS} che porta la caratteristica $V_{GS} = 0$ al centro del piano I – V; in particolare, quando $V_{GS} = 0$ il MOSFET a svuotamento si comporta come un MOSFET ad arricchimento al quale è applicata una tensione di over – drive $V_{GS} - V_{th} = V_{TD}$. Di conseguenza, la rappresentazione analitica di un tale dispositivo è la seguente:

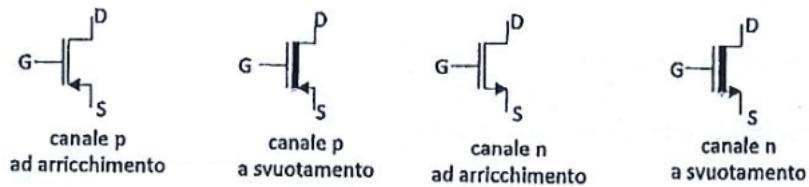
$$\begin{cases} V_{GS} < -|V_{TD}| \Leftrightarrow I_D = 0 \\ V_{GS} > -|V_{TD}| \Leftrightarrow \begin{cases} I_D = k[2(V_{GS} + |V_{TD}|)V_{DS} - V_{DS}^2] \Leftrightarrow V_{DS} < V_{GS} + |V_{TD}| \\ I_D = k(V_{GS} + |V_{TD}|)^2 \Leftrightarrow V_{DS} \geq V_{GS} + |V_{TD}| \end{cases} \end{cases}$$

Per $V_{GS} = -V_{TD}$ la corrente si annulla, ed è diversa da zero per $V_{GS} = 0$.

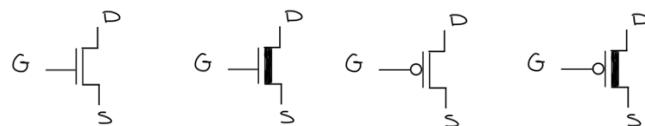
Esistendo sia il MOSFET a svuotamento con canale N e con canale P, è possibile individuare quattro tipologie di MOSFET:

- MOSFET ad **arricchimento** con canale N;
- MOSFET a **svuotamento** con canale N;
- MOSFET ad **arricchimento** con canale P;
- MOSFET a **svuotamento** con canale P.

I simboli circuituali che rappresentano questi dispositivi sono i seguenti:

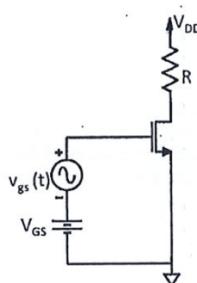


Questa notazione **utilizza la stessa convenzione fatta per il BJT** per la quale la freccia indica il terminale di Source; tuttavia, nei MOSFET i terminali di Drain e Source sono realmente simmetrici, non essendoci alcuna differenza tra i drogaggi, ed invertendoli il funzionamento del dispositivo non è alterato (un BJT non funzionerebbe). Quindi, la freccia che indica il Source è ridondante, servirebbe solo la sua direzione per comunicare il fatto che si parla di MOSFET a canale P o N; invece, al posto di introdurre una nomenclatura così difficile per un dettaglio così piccolo, si può sfruttare la funzione di dispositivi di negazione che i MOSFET a canale P assumono in elettronica digitale per utilizzare una nomenclatura più semplice:



MODELLO A PICCOLO SEGNALE DEL MOSFET

Definire un punto di funzionamento per un circuito che fa uso di un MOSFET non è tanto diverso rispetto a quanto fatto finora con i BJT, se non per la semplificazione dovuta al fatto che $I_G = 0$ (in un BJT $I_B \neq 0$ e andava messa in considerazione nella polarizzazione, producendo dei calcoli scomodi). Per studiare il comportamento amplificativo di un MOSFET è necessario studiare dapprima il **comportamento del dispositivo in corrispondenza di segnali variabili**, osservando quali sono le condizioni per la linearizzazione del sistema.



Nel circuito proposto, **in ingresso al MOSFET è applicata la tensione:**

$$v_{GS}(t) = v_{gs}(t) + V_{GS}$$

Con $v_{gs}(t)$ **forma d'onda sinusoidale.** È possibile ricavare la seguente equazione:

$$\begin{aligned} i_D(t) &= k(v_{GS}(t) - V_{th})^2 = k(v_{gs}(t) + V_{GS} - V_{th})^2 \\ &= k(V_{GS} - V_{th})^2 + 2k(V_{GS} - V_{th})v_{gs}(t) + k(v_{gs}(t))^2 \end{aligned}$$

Essa permette di rilevare un legame quadratico tra la corrente che circola nel Drain e il segnale da amplificare; tuttavia, **l'amplificazione necessita un legame lineare,** che può essere ottenuto considerando che:

$$k(v_{gs}(t))^2 \ll 2k(V_{GS} - V_{th})v_{gs}(t)$$

Ovvero:

$$v_{gs}(t) \ll 2(V_{GS} - V_{th})$$

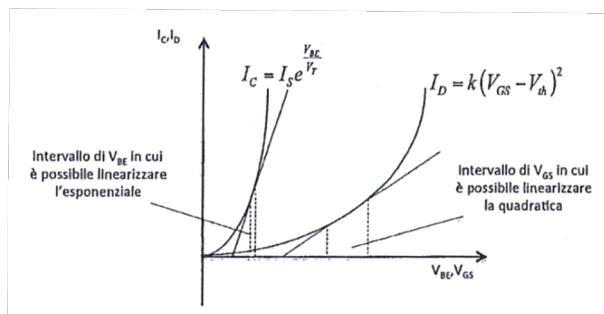
Allora:

$$i_D(t) = k(V_{GS} - V_{th})^2 + 2k(V_{GS} - V_{th})v_{gs}(t) = I_D^* + g_m v_{gs}(t)$$

Il legame, sotto le precedenti ipotesi di approssimazione, è lineare e il dispositivo si comporta da generatore controllato; I_D^* è detta **corrente statica** e rappresenta la **corrente che circola nel Drain quando il segnale sinusoidale non è applicato.**

L'ipotesi grazie alla quale è possibile fare l'approssimazione in esame è detta, anche in questo caso, **ipotesi di piccolo segnale** ed il modello costruito a partire da essa è detto **modello a piccolo segnale.**

Prima di procedere all'illustrazione del modello a piccolo segnale, sono necessarie alcune considerazioni. **Il legame tra la corrente da controllare e il segnale da amplificare, in un BJT è esponenziale e in un MOSFET è quadratico;** quindi, **in quest'ultimo la condizione di piccolo segnale sarà più rilassata e permissiva** (di circa 100 volte), **con la possibilità di amplificare una quantità maggiore di segnali con lo stesso circuito amplificativo:**



Questo vantaggio, da solo, **garantirebbe l'annichilimento dello studio dei BJT;** tuttavia, li si studia ancora perché non è tutto oro ciò che luccica e **il MOSFET porta con sé dei problemi.** In particolare, **il guadagno del MOSFET non è mai così alto come quello dei BJT:** in questi ultimi, g_m è determinata a partire da un rapporto il cui denominatore è una quantità estremamente piccola,

mentre in un MOSFET è governato da k , che assume valori più contenuti. Quindi, in linea di principio, un BJT amplifica meno segnali ma con un maggior guadagno, mentre un MOSFET amplifica una banda maggiore di segnali ma con un guadagno minore.

Si vogliono confrontare, per un BJT ed un MOSFET, le definizioni di g_m , rilevando in particolare la differenza negli ordini di grandezza:

$$g_{mBJT} = \frac{i_c(t)}{V_t} = \frac{i_c(t)}{25 \cdot 10^{-3} V} \wedge g_{mMOS} = 2k(V_{GS} - V_{th}) \approx 4k$$

Dal momento in cui **k non assume valori eccessivamente grandi**, il fatto che nel parametro del BJT si trova una quantità molto piccola al denominatore implica quasi automaticamente che il guadagno di un amplificatore a BJT è maggiore del guadagno di un amplificatore a MOSFET:

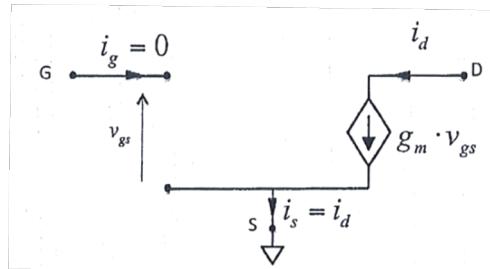
$$g_{mBJT} > g_{mMOS}$$

Infine, si procede a fornire una **definizione alternativa di guadagno per il MOSFET**:

$$g_m = 2k(V_{GS} - V_{th}) = 2\sqrt{k} \cdot \sqrt{k}(V_{GS} - V_{th}) = 2\sqrt{kI_D^*}$$

Quindi, sotto le ipotesi di piccolo segnale, la corrente di Drain può essere distinta in una componente dovuta alla polarizzazione, I_D^* , e in una che dipende dal segnale, $i_d(t) = g_m v_{gs}(t)$. Analizzando solo dal punto di vista dinamico l'amplificatore a MOSFET, è possibile ricavare un **modello semplificato lineare**, che prende il nome di **modello a piccolo segnale**, descritto dalle seguenti equazioni:

$$\begin{cases} |i_s(t)| = |i_d(t)| = |g_m v_{gs}(t)| \\ i_g(t) = 0 \end{cases}$$

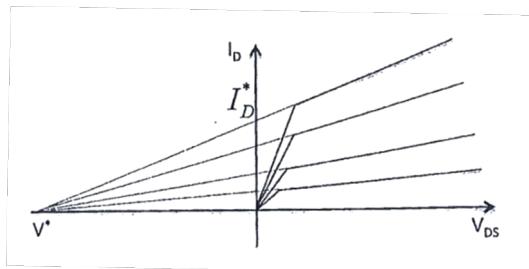


Per assicurarsi che la corrente di Gate sia nulla è necessario supporre una resistenza r_π infinita. La resistenza in questione permette anche di far cadere tutto il segnale $v_{gs}(t)$ sul generatore controllato (non essendoci partitori di tensione), rendendo non necessario qualsiasi doppio stadio o buffer di tensione. In realtà la corrente $i_g(t)$ non è nulla, esiste un piccolissimo contributo (detto corrente di leakage) ma che è ingegneristicamente irrilevante.

Quando sono state enunciate le caratteristiche del MOSFET è stata mostrata l'**indipendenza della corrente di drain I_D dalla tensione V_{DS}** ; in realtà, in maniera del tutto analoga a come accade nel BJT, la retta che rappresenta tali caratteristiche non è perfettamente orizzontale ma presenta una leggerissima inclinazione, la cui causa va ricercata nell'effetto di modulazione della lunghezza di canale (nel BJT prendeva il nome di effetto di modulazione dello spessore di base, o effetto Early).

Il motivo della teorica indipendenza era attribuito al fatto che **il punto in cui il canale si strozza** (quando viene applicata una tensione superiore a quella di pinch-off), x^* , resta **sempre molto prossimo al Drain**, in modo da avere una lunghezza di canale L ed una resistenza R costanti; nella realtà **si riscontra un lievissimo spostamento del punto x^*** , che si allontana dal Drain accorciando la regione di carica, con le conseguenti **diminuzione della resistenza R e dipendenza dalla V_{DS}** . Analiticamente, l'effetto in questione viene modellato dalla seguente equazione:

$$I_D = k(V_{GS} - V_{th})^2(1 + \lambda V_{DS})$$

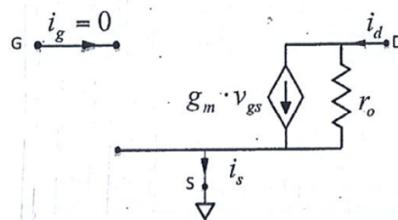


Dove $\lambda = [V^{-1}]$ è **un parametro empirico dipendente dalla tecnologia di fabbricazione del MOSFET**. Si noti che i prolungamenti delle caratteristiche a V_{GS} costante intersecano l'asse delle tensioni nel punto V^* , rilevato essere:

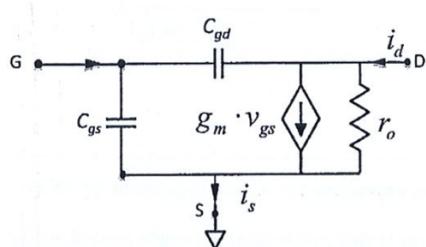
$$V^* = -\frac{1}{\lambda}$$

Il parametro λ svolge lo stesso ruolo della tensione V_A dell'effetto Early: entrambi servono a **modellare un effetto di linearità lievissimo** e, pertanto, devono assumere dei valori limiti (∞ per V_A e 0 per λ) affinché siano irrilevanti. Volendo includere l'effetto di modulazione della lunghezza di canale nel modello a piccolo segnale, andrebbe **inserito in parallelo al generatore controllato una resistenza di uscita r_o** pari a:

$$r_o = \frac{1}{\lambda I_D}$$



Come nel BJT, anche nel MOSFET sono presenti delle capacità parassite (legate alla presenza del condensatore MOS) che manifestano la loro presenza solo alle alte frequenze; esse possono essere schematizzate come una **capacità tra Source e Gate, C_{gs}** , ed una **capacità tra Drain e Gate, C_{gd}** :



Poiché il modello a piccolo segnale è ricavato a partire da un punto di funzionamento statico posto in regione di pinch – off, il canale conduttivo tra Source e Drain è sicuramente strozzato e le due capacità non possono avere lo stesso valore. In particolare, la variazione di carica associata alla variazione della tensione $v_{gs}(t)$ (quindi C_{gs}) è maggiore della corrispondente variazione di carica associata alla variazione della tensione $v_{gd}(t)$ (quindi C_{gd}), dal momento in cui nel Drain è presente una regione svuotata con poca disponibilità di carica; numericamente:

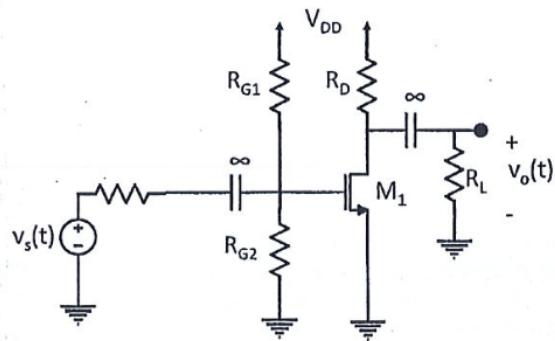
$$C_{gs} \approx 3C_{gd}$$

Con valori specifici che variano tra i pF per la prima capacità e frazioni di pF per la seconda. Anche per i MOSFET queste capacità parassite determinano le frequenze di taglio superiore in cui un eventuale amplificatore a MOSFET amplifica; tuttavia, rispetto al corrispettivo del BJT, è nettamente più basso, rendendo il MOSFET non adatto all'amplificazione di segnali ad alte frequenze.

AMPLIFICATORI ELEMENTARI A MOSFET

Il modello a piccolo segnale mostrato in precedenza presenta una notevole somiglianza con l'omonimo del BJT, ad eccezione del fatto che per il MOSFET la resistenza r_π può essere considerata infinita; sotto queste assunzioni, alle medie frequenze i due modelli equivalenti sono identici. La comodità di questa identità sta nel fatto che i parametri da ricavare nel modello del MOSFET possono essere ottenuti come se si fosse in un BJT senza r_π , semplificando di molto la polarizzazione di eventuali circuiti amplificatori.

Un circuito amplificatore a MOSFET non differisce molto, almeno topologicamente, da uno a BJT, se non per il fatto che in comune tra maglia di ingresso e di uscita è il terminale di Source di un MOSFET, il cui nodo è dinamicamente a massa, collegando tra Gate e massa il segnale di ingresso e tra Drain e massa il carico. Il circuito si configura come segue:



Il circuito in questione è studiato (senza pretendere di fare alcuna digressione sulla polarizzazione e sulla scelta delle resistenze) sotto le ipotesi di piccolo segnale, in modo tale da poter essere tradotto agilmente in una versione lineare equivalente, almeno fintantoché la frequenza del segnale non diventa sufficientemente elevata da rendere visibili gli effetti delle capacità interne del dispositivo attivo.

Volendo individuare il guadagno di tensione a piccolo segnale, si considera che $v_o(t)$ è determinata dal fluire della corrente $g_m v_{gs}(t)$ nel parallelo delle tre resistenze incontrate tra Drain e massa:

$$v_o(t) = -g_m v_{gs}(t)(r_o || R_D || R_L)$$

Osservando la maglia di ingresso, **si può notare che $v_{gs}(t)$** , dato l'elevato valore delle resistenze R_{G1} e R_{G2} , **praticamente coincide con la tensione di ingresso $v_s(t)$** . Mettendo insieme tutti i risultati ottenuti:

$$A_v = \frac{v_o(t)}{v_s(t)} = -\frac{g_m v_s(t)(r_o || R_D || R_L)}{v_s(t)} = -g_m(r_o || R_D || R_L)$$

Volendo valutare la resistenza di uscita, **si consideri che la resistenza di ingresso coincide con il parallelo tra R_{G1} e R_{G2}** e, come nel caso dell'amplificatore ad emettitore comune, dall'uscita si può determinare:

$$R_{out} = R_D || r_o$$

In questa sede **l'amplificatore a MOSFET non sarà utilizzato molto** ed il motivo sarà chiaro a breve, dopo un'opportuna comparazione con l'omonimo a BJT.

Semplicità di progettazione	MOSFET (più semplice)
Linearità	MOSFET (condizioni più rilassate)
R_{in}	MOSFET (più elevata)
Frequenza	BJT (Banda più larga e ω_H più elevata)
Guadagno di amplificazione	BJT (più elevata)

La semplicità di progettazione del MOSFET non risiede solo nella comodità del suo modello lineare (in cui è assente r_π) **ma anche nel fatto che la corrente di Gate è nulla**, di conseguenza non è necessario impostare determinati valori di resistenza per farvi scorrere corrente (come accade nel nodo di base con un BJT), ottenendo **un grado di libertà in più**.

L'ELETTRONICA INTEGRATA E L'AMPLIFICATORE DIFFERENZIALE

Fino a questo momento **non si è detto nulla sulla realizzazione pratica dei circuiti** (amplificatori, raddrizzatori, ecc ...) **ma solo sui dispositivi in essi contenuti** (diodi, BJT, MOSFET, ecc ...). **Si può pensare che nei devices di uso quotidiano** (smartphone, computer, ecc ...) **si trovino** pedissequamente le resistenze, i condensatori, gli induttori e i dispositivi **così come sono stati** schematizzati nel corso della trattazione; **la realtà non è questa**, sebbene ci si possa trovare davanti ad oggetti di questo tipo. Per comprendere meglio quanto si sta dicendo, **si rende necessaria la distinzione tra elettronica discreta e integrata**:

- **Elettronica discreta**, è quella branca dell'elettronica che tratta ogni componente in quanto un oggetto fisico a sé stante e da aggiungere al circuito in un secondo momento rispetto alla realizzazione;

- **Elettronica integrata**, è quella branca dell'elettronica che tratta un circuito non come composto da diversi componenti a sé stanti ma come una serie di effetti fisici che si concatenano su uno stesso dispositivo, realizzato integralmente.

In **elettronica discreta**, un **amplificatore a BJT** verrà realizzato collegando quattro resistenze, quattro capacità e un BJT su una stessa scheda, mentre in **elettronica integrata** verrà realizzato realizzando su uno stesso wafer di silicio delle **regioni ad effetto resistivo**, delle **regioni ad effetto capacitivo** ed una ad effetto transistore in un ordine tale da restituire il comportamento amplificativo, per poi collegare i terminali di ingresso ed uscita con l'esterno.

Ovviamente, l'**elettronica discreta** è più modulare ma richiede una quantità di spazio notevolmente maggiore, mentre l'**elettronica integrata** permette di condensare diversi dispositivi in uno spazio microscopico ma al costo della modularità, che è impensabile; di conseguenza, per la realizzazione di un dispositivo in **elettronica integrata**, è necessaria una accurata progettazione ed un accurato testing del prototipo, dal momento in cui eventuali errori non possono essere risolti semplicemente scambiando il componente "difettoso". Nel tempo, per le applicazioni commerciali è stato preferito un approccio più aggressivo nei confronti dell'**elettronica integrata**, che ha favorito un rapido sviluppo del processo di miniaturizzazione, con il quale oggi è possibile avere tra le mani dispositivi molto complessi in uno spazio notevolmente ridotto (negli smartphone la maggior parte dello spazio è occupata dalle batterie).

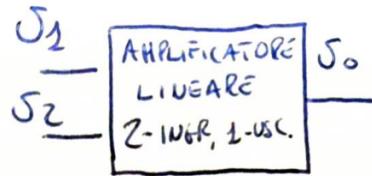
Nella vita quotidiana capita spesso di incontrare dei **dispositivi i cui cavi di alimentazione o di input/output sono caratterizzati da due morsetti di ingresso**:



Su uno dei due morsetti va il segnale da amplificare, proveniente dall'esterno e all'interno del quale possono essere presenti rumori che non contengono l'informazione desiderata ma che sporcano solo il segnale, mentre **sull'altro morsetto va qualsiasi rumore non informativo che proviene dall'esterno**. **Amplificare solo il segnale sul primo morsetto non produce una corretta amplificazione perché si andrebbe a lavorare anche sulle componenti indesiderate**; si rende necessario **un tipo di amplificatore che agisca sulle differenze**, in modo da prima eliminare il rumore e poi amplificare solo il segnale informativo.

Nei modelli di amplificatore mostrati finora **sono stati accennati dei filtri particolari**, che agiscono però solo sulle componenti continue; infatti, **i capacitori di disaccoppiamento si occupano di eliminare dall'amplificazione del segnale dinamico le componenti continue della tensione in ingresso**. Tuttavia, questo meccanismo non è sufficiente, sia perché il **rumore non è necessariamente rappresentato da una tensione continua e sia perché in elettronica integrata è molto difficile realizzare componenti passivi** (resistori, capacitori e induttori) di valore considerevole (il loro utilizzo va limitato al minimo). **Si rende necessario un amplificatore che preleva il segnale sui due terminali, elimina le interferenze comuni ed amplifica solo il segnale manipolato, senza adoperare capacitori**; il circuito elettronico responsabile di questo tipo di amplificazione è detto **amplificatore differenziale**.

Prima di procedere alla descrizione precisa di questo dispositivo, sono necessarie delle considerazioni sulla linearità di un sistema a due ingressi ed un uscita (MISO):



v_1 e v_2 sono gli ingressi e v_o l'uscita. La relazione tra ingresso ed uscita è descritta dalla seguente combinazione lineare:

$$v_o = A_1 v_1 + A_2 v_2$$

Per la quale è possibile individuare le seguenti grandezze:

$$A_d = \frac{A_1 - A_2}{2} \wedge A_{cm} = A_1 + A_2$$

La stessa relazione ingresso – uscita può essere espressa in funzione di queste grandezze:

$$v_o = A_d(v_1 - v_2) + A_{cm} \frac{v_1 + v_2}{2}$$

Infatti:

$$\begin{aligned} v_o &= \frac{A_1 - A_2}{2}(v_1 - v_2) + (A_1 + A_2) \frac{v_1 + v_2}{2} \\ &= \frac{1}{2}[A_1 v_1 - A_2 v_1 - A_1 v_2 + A_2 v_2 + A_1 v_1 + A_2 v_1 + A_1 v_2 + A_2 v_2] \\ &= \frac{1}{2}[2A_1 v_1 + 2A_2 v_2] = A_1 v_1 + A_2 v_2 \end{aligned}$$

Definendo $v_d = v_1 - v_2$ e $v_{cm} = v_1 + v_2$, il sistema può essere descritto dalla seguente equazione:

$$v_o = A_d v_d + A_{cm} v_{cm}$$

A_d è detto **guadagno differenziale** e A_{cm} **guadagno di modo comune** ed agiscono, rispettivamente, sul segnale differenziale v_d e sul segnale di modo comune v_{cm} . Si può ragionevolmente pensare che v_d corrisponda al segnale puro da amplificare, dal momento in cui su uno dei due terminali del cavo è applicato segnale + rumore e sull'altro solo rumore:

$$v_d = (\text{segnale} + \text{rumore}) - (\text{rumore}) = \text{segnale}$$

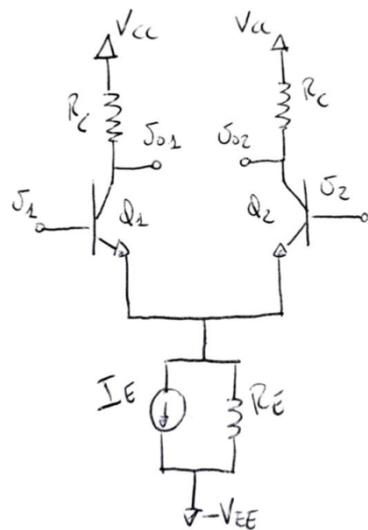
Per quanto riguarda il **segnale di modo comune**, esso rappresenta il **valor medio dei segnali in ingresso**. Se ai due segnali di ingresso fossero sovrapposte una componente in continua (cioè se vi fosse posto un generatore di tensione continua in serie), il segnale differenziale non ne sarebbe condizionato mentre il segnale di modo comune gli corrisponderebbe.

Si definisce **rapporto di reiezione di modo comune** (CMRR, Common Mode Rejection Ratio) come:

$$\text{CMRR} = \frac{A_d}{A_{cm}}$$

Si misura, talvolta, in decibel e mette in relazione quanto il sistema a due ingressi filtra le componenti comuni tra i due segnali e quanto, allo stesso tempo, amplifica la loro differenza. Più il CMRR è alto e più l'amplificatore si sta comportando “bene” in quanto amplificatore differenziale.

Sarebbe molto utile pensare ad un dispositivo che ha un **guadagno differenziale molto alto** ed un **guadagno di modo comune molto basso**; un circuito di questo tipo esiste ed è detto **amplificatore differenziale**:



Si possono osservare **due BJT** (esiste anche la versione con i MOSFET) che condividono l'emettitore, ai capi di ogni base sono collegati i segnali v_1 e v_2 e si può notare l'assenza di **capacitori di accoppiamento** (le componenti continue sono filtrate dal basso A_{cm}). I **due transistor** di uno stesso amplificatore differenziale **devono essere identici (matched)**, ovvero **realizzati in tecnologia integrata uno accanto all'altro in modo da poter avere lo stesso valore β** ed esser soggetti alla stessa temperatura; si può fare lo stesso discorso per le resistenze R_C (che spesso sono altri transistor in configurazione tale da mostrare un comportamento resistivo).

Ogni ramo (o gamba/leg in gergo) **rappresenta un amplificatore ad emettitore comune**, i cui emettitori sono condivisi; **la rete di polarizzazione non viene adoperata in nessuno dei due amplificatori perché la polarizzazione viene delegata ad un generatore di corrente** (reale perché c’è la resistenza in parallelo R_E , che non è un resistore ma un effetto resistivo del generatore reale), mentre **l’uscita viene prelevata come differenza di potenziale tra v_{o1} e v_{o2}** in modo da **non risentire di una componente continua dettata dalla polarizzazione stessa** e da **non necessitare dei capacitori sui collettori**.

Considerando $v_o = v_{o1} - v_{o2}$, il circuito può essere modellato dalla seguente relazione ingresso – uscita:

$$v_o = A_d v_d + A_{cm} v_{cm} = (-g_m R_C) v_d + \left(-\frac{R_C}{R_E} \right) v_{cm}$$

Con:

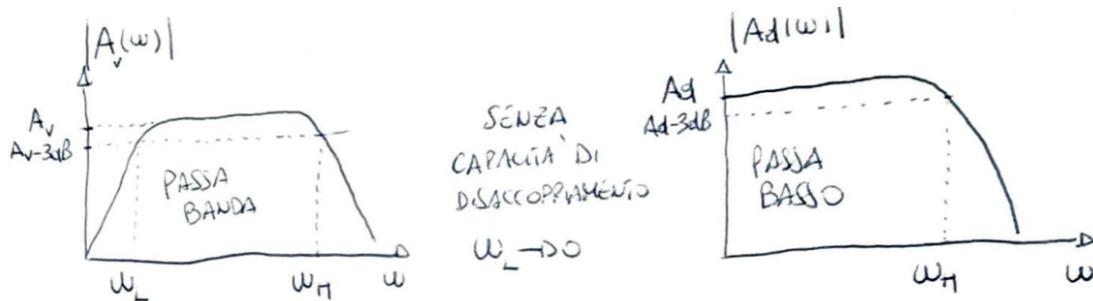
$$g_m = \alpha \frac{I_E}{V_T}$$

La resistenza R_E , seppur non infinita, è di diversi ordini di grandezza superiore a R_C e ciò permette all'amplificatore di restituire l'effetto desiderato, filtrando solo la componente di modo comune; infatti, si può osservare che il CMRR è abbastanza elevato:

$$\text{CMRR} = g_m \cdot R_E$$

In corrispondenza di un **generatore di corrente ideale**, quindi con $R_E = \infty$, l'amplificatore differenziale agirebbe nelle sue condizioni ideali, essendo **CMRR = ∞** .

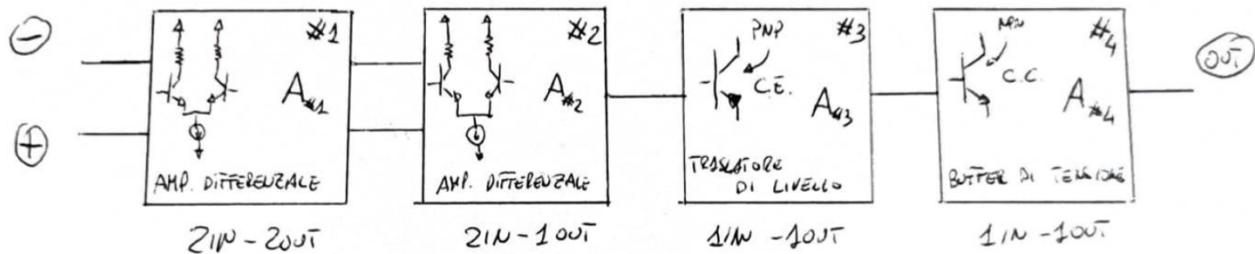
Poiché di un amplificatore differenziale interessa prettamente l'amplificazione A_d , ci si sofferma solo sulla sua **risposta in frequenza** trascurando quella di A_{cm} . Sugli ingressi non ci sono capacitori, visto che il filtraggio delle componenti continue avviene grazie a A_{cm} , mentre le singole uscite v_{o1} e v_{o2} possono avere una **componente continua** (la caduta di tensione su R_C) ma, essendo l'uscita v_o la loro differenza, essa viene annichilita. L'assenza di capacitori nel circuito fa sì che non ci sia frequenza di taglio inferiore che limita la banda passante; l'amplificazione si comporta, quindi, come un **filtro passa basso** di guadagno A_d , con la sola frequenza di taglio superiore ω_H determinata dagli effetti capacitivi delle giunzioni nel BJT.



Sulla sinistra è schematizzata la **risposta in frequenza di un amplificatore a BJT "classico"**, mentre sulla destra il suo equivalente differenziale; i due comportamenti sono diversi proprio per l'assenza dei capacitori di disaccoppiamento nel differenziale, che **annulla la frequenza di taglio inferiore** e permette un **comportamento da filtro passa basso** (a differenza del bassa banda di un amplificatore a BJT).

L'AMPLIFICATORE OPERAZIONALE

L'amplificatore **operazionale** (o solo operazionale) è un **componente fondamentale** per l'elettronica, analogica e digitale, perché, come suggerisce il nome, permette di effettuare tutta una serie di operazioni non possibili con gli altri amplificatori affrontati finora. Per realizzare un operazionale (non addentrandosi nei calcoli specifici in quanto non di interesse in questa sede) sono necessari **4 stadi amplificatori connessi in serie**:



- **Stadio #1 (2 ingressi e 2 uscite)**

È un amplificatore differenziale collegato ai due morsetti di ingresso dell'operazionale; per sua natura amplifica solo la differenza degli ingressi e il suo guadagno $A_{\#1}$ è sicuramente molto maggiore di $1 V/V$. La priorità di questo stadio non è il guadagno elevato, bensì un'elevata resistenza di ingresso, proporzionale a $r_\pi = \beta/g_m$; quindi, di proposito si riduce g_m (ovvero la corrente I_C), perdendo sul guadagno (pari a $g_m R_C$) ma guadagnando su R_{in} .

In realtà la resistenza di ingresso sarebbe pari a $r_\pi || R_B$, visto che l'operazionale è composto da due amplificatori CE; tuttavia, non essendo polarizzato con le resistenze sulla base ma con il generatore di corrente sull'emettitore, R_B è assente e R_{in} dipende solo da r_π . Analogamente, il guadagno (in modulo) sarebbe $g_m \cdot R_C || R_L$ ma il carico coincide con la resistenza di ingresso del secondo stadio, il cui valore è solitamente più elevato di quello di R_C , potendo essere escluso.

- **Stadio #2 (2 ingressi e 1 uscita)**

Anche questo stadio è un amplificatore differenziale, che prende in ingresso i segnali sul primo stadio. Non è necessaria una resistenza R_{in} elevata (rispetto allo stadio #1) perché bisogna spingere di più sul guadagno; quindi, risulterà $A_{\#2} > A_{\#1}$ e rappresenterà il contributo principale del guadagno complessivo dell'operazionale. L'uscita è prelevata solo da una delle due leg del differenziale (questo stadio è detto differenziale single – ended) e, pertanto, ci sarà il segnale con una componente continua sovrapposta (la caduta di tensione sulla resistenza R_C)

- **Stadio #3 (1 ingresso e 1 uscita)**

Questo stadio è un traslatore di livello che si occupa di eliminare la componente continua proveniente dallo stadio precedente. Oltre a sapere che fa utilizzo di un BJT PNP (o di un PMOS) e che il suo guadagno $A_{\#3}$ è sicuramente minore di $A_{\#1}$ ma confrontabile con $A_{\#1}$, in questa sede non è necessario sapere null'altro su questo stadio.

- **Stadio #4 (1 ingresso e 1 uscita)**

L'uscita di questo stadio coincide con l'uscita dell'operazionale ma lo stadio in sé non amplifica, bensì attenua con un guadagno $A_{\#4}$ di poco inferiore a 1 (circa il 95%). È un buffer di tensione realizzato come amplificatore a collettore comune ed ha come principale fine quello di mostrare una resistenza di uscita bassissima (il più delle volte trascurabile).

I valori ragionevoli per i singoli guadagni di ogni stadio possono essere:

$$A_{\#1} = 70 \frac{V}{V} \wedge A_{\#2} = 300 \frac{V}{V} \wedge A_{\#3} = 50 \frac{V}{V} \wedge A_{\#4} = 0.95 \frac{V}{V}$$

Il fattore di amplificazione complessivo A_o dell'operazionale è dato dal prodotto dei guadagni dei singoli stadi (perché sono in serie):

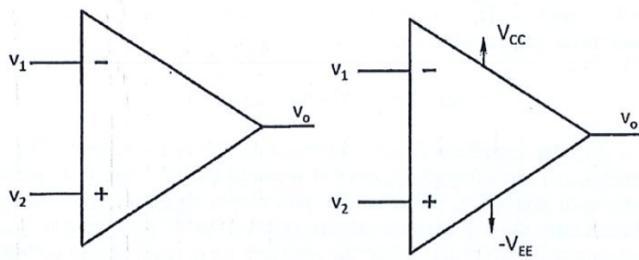
$$A_o = (70 \cdot 300 \cdot 50 \cdot 0.95) \frac{V}{V} = 997.5 \frac{kV}{V}$$

Ovvero **un milione V/V!** L'elevatissimo guadagno dell'operazionale viene affiancato a:

1. **Un'enorme resistenza di ingresso** (specialmente negli operazionali a MOSFET);
2. **Una bassissima resistenza di uscita**, grazie al buffer di tensione sul quarto stadio;
3. **Un elevatissimo CMRR**.

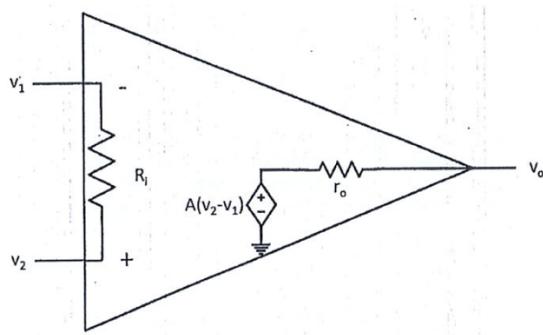
Per la banda passante (per ora) ci si limita a dire che il primo stadio differenziale fa sì che non ci sia frequenza di taglio inferiore, mentre per la frequenza di taglio superiore si rimanda la discussione ad un secondo momento.

Il **simbolo circuitale** con cui si riferisce un amplificatore operazionale è il seguente e permette di notare come esso sia **un componente** (non è un dispositivo) a **cinque terminali**, di cui due di alimentazione:



I due simboli sono equivalenti, sebbene si preferisca il primo per la sua semplicità di rappresentazione. Il + e il - dei morsetti di ingresso stanno a significare che il segnale di uscita è **in fase o meno con il morsetto** e vengono detti, rispettivamente, **morsetto non invertente** e **morsetto invertente** per un motivo che sarà più chiaro a breve.

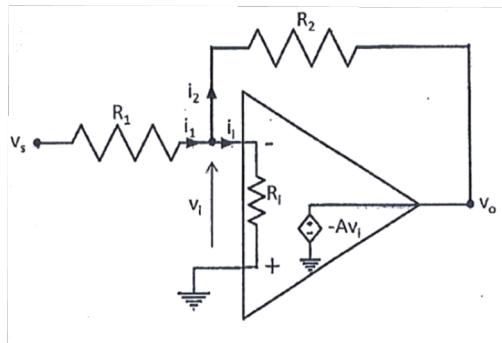
Per analizzare un eventuale modello equivalente di un amplificatore operazionale **si può ricorrere alla seguente schematizzazione**:



Sebbene un **operazionale possa essere preso singolarmente e usato come amplificatore differenziale** (detto amplificatore a ciclo aperto), in questa sede non verrà sfruttato il suo alto guadagno e sarà spesso inserito in una rete che mette in comunicazione il terminale di uscita con uno dei due terminali di ingresso (e sarà usato come **amplificatore a ciclo chiuso**, o reazionario); una rete di questo tipo è detta **rete di reazione** e il collegamento è definito **collegamento in retroazione**.

Il collegamento in retroazione prende anche il nome di feedback e si distingue in feedback positivo (quando la derivata della grandezza è positiva) e **feedback negativo** (quando la derivata della grandezza è negativa). Il motivo per cui **la retroazione di un circuito ad operazionale è negativo** è dovuto al fatto che **il feedback positivo conduce molto spesso a situazioni di instabilità, mentre quello negativo si stabilizza.**

Nella seguente figura è mostrata una delle possibili configurazioni di una rete di reazione, in cui l'operazionale è solo una parte del circuito e il cui guadagno non andrà a condizionare minimamente il guadagno complessivo; si noti che il segnale di ingresso non è applicato direttamente ad uno dei terminali dell'operazionale ma alla resistenza R_1 . La configurazione seguente è detta **configurazione invertente**:



L'obiettivo di questa prima trattazione è il **calcolo del guadagno di tensione**:

$$A_v = \frac{v_o(t)}{v_s(t)}$$

Si suppone di conoscere il guadagno a ciclo aperto (ovvero il guadagno dell'operazionale A_o) e le resistenze R_{in} e r_o ; nei componenti reali, quest'ultima assume valori così bassi da poter essere trascurata, ovvero si può considerare nulla la caduta di tensione su di essa (infatti non è stata inserita nella figura). Con riferimento alla schematizzazione in figura, si può sicuramente dire che:

$$v_o(t) = -A_o \cdot v_i(t)$$

Il cui segno meno deriva dal fatto che la tensione $v_i(t)$ è assunta positiva sul terminale invertente. Dal circuito è possibile ricavare facilmente l'espressione delle tre correnti:

$$i_1(t) = \frac{v_s(t) - v_i(t)}{R_1}$$

$$i_2(t) = \frac{v_i(t) - v_o(t)}{R_2}$$

$$i_i(t) = \frac{v_i(t)}{R_i}$$

In modo che la LKC sia verificata:

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

Per cui:

$$\frac{v_s(t) - v_i(t)}{R_1} = \frac{v_i(t)}{R_i} + \frac{v_i(t) - v_o(t)}{R_2}$$

Sostituendo con l'espressione di $v_i(t)$ in funzione di $v_o(t)$ precedentemente introdotta:

$$\frac{v_s(t)}{R_1} = -v_o(t) \left[\frac{1}{A_o} \left(\frac{1}{R_i} + \frac{1}{R_1} + \frac{1}{R_2} \right) + \frac{1}{R_2} \right] = -\frac{v_o(t)}{R_2} \left[\frac{1}{A_o} \left(\frac{R_2}{R_i} + \frac{R_2}{R_1} + 1 \right) + 1 \right]$$

Da cui, infine:

$$A_v = \frac{v_o(t)}{v_s(t)} = -\frac{R_2}{R_1} \cdot \frac{1}{1 + \frac{1}{A_o} \left(1 + \frac{R_2}{R_1} + \frac{R_2}{R_i} \right)} \approx -\frac{R_2}{R_1}$$

L'approssimazione in chiusura è dovuta al fatto che:

$$A_o \gg \left(1 + \frac{R_2}{R_1} + \frac{R_2}{R_i} \right)$$

E quindi:

$$\frac{1}{A_o} \left(1 + \frac{R_2}{R_1} + \frac{R_2}{R_i} \right) \approx 0$$

Il risultato appena rilevato suggerisce che **il guadagno di un amplificatore ad operazionale (diverso dall'amplificatore operazionale) non dipende dal guadagno di quest'ultimo ma unicamente dal rapporto delle due resistenze che compongono la rete di reazione.** Rispetto al guadagno di altri dispositivi, questo risultato si porta dietro numerosi vantaggi in termini di predicitività e stabilità del risultato rispetto alla variazione dei parametri dei componenti elettronici che compongono l'operazionale.

Alla base di questa semplificazione si trova il valore elevato di A_o , che può essere supposto infinito; infatti, quando si riscontrano discrepanze tra il valore di amplificazione A_v teorico e reale non sono mai dovute a A_o ma all'incertezza sui valori assunti dai resistori R_1 e R_2 .

In questo modo si verifica, ai capi della resistenza R_i , il fenomeno di **cortocircuito ideale**:

$$v_i(t) = -\frac{v_o(t)}{A \rightarrow \infty} \rightarrow 0$$

Ovviamente, nella pratica $v_i(t)$ non è nullo ma circa un milionesimo di $v_o(t)$, che è una frazione ingegneristicamente nulla. Osservando la corrente che scorre ai capi di R_i :

$$i_i(t) = \frac{v_i(t)}{R_i}$$

Si può dedurre che **il cortocircuito virtuale impone una corrente entrante nei terminali dell'operazionale nulla** e la LKC può essere aggiornata:

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

Ma si aggiornano anche i valori delle altre correnti:

$$i_1(t) = \frac{v_s(t)}{R_1}$$

$$i_2(t) = \frac{-v_o(t)}{R_2}$$

Questo risultato avvalora quello sul guadagno:

$$v_o(t) = -R_2 i_2(t) \wedge v_s(t) = R_1 i_1(t) \Rightarrow A_v = \frac{v_o(t)}{v_s(t)} = -\frac{R_2 i_2(t)}{R_1 i_1(t)} = -\frac{R_2 i_1(t)}{R_1 i_1(t)} = -\frac{R_2}{R_1}$$

Un altro modo per **giustificare questi risultati** è considerare **infinita la resistenza di ingresso R_{in}** (che in realtà ha solo valori molto elevati, da qualche $k\Omega$ a $M\Omega$).

Le ipotesi di guadagno operazionale e resistenza di ingresso infinita sono **ideali nonostante i valori reali possano consentire una buona approssimazione**. Confrontando i valori reali e quelli ideali dei parametri che determinano il comportamento di un amplificatore ad operazionale

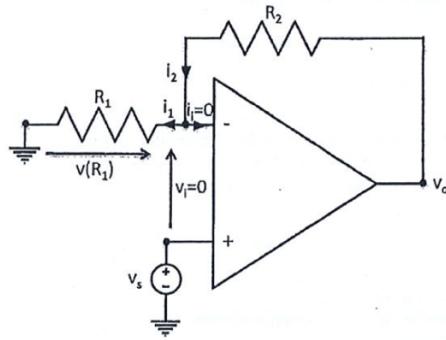
Parametro	Valore reale	Valore ideale
A_o	$10^6 V/V$	∞
R_{in}	$10^6 \Omega$	∞
r_o	1Ω	0
CMRR	Alto	∞
Banda	ω_L nulla e ω_H alta	∞

In genere **si lavorerà con i valori ideali**, tenendo comunque in considerazione che **gli stessi risultati sono ingegneristicamente equivalenti se ottenuti con i valori reali**.

Di seguito sono proposte alcune delle configurazioni fondamentali dell'amplificatore ad operazionale:

- **Amplificatore ad operazionale in configurazione non invertente**

Nell'amplificatore ad operazionale invertente **il guadagno è negativo**, ciò significa che **il segnale è amplificato ed invertito**; lo si poteva intuire anche considerando **il nome del terminale su cui il segnale stesso è applicato nell'operazionale**, terminale invertente. Per non invertire il segnale bisogna considerare una configurazione in cui **il guadagno non sia negativo**; si può intuire facilmente che **il segnale andrà applicato al terminale non invertente dell'operazionale e la resistenza R_1 direttamente a massa**:



Sfruttando il cortocircuito virtuale, si può dire che la tensione di ingresso $v_s(t)$ si ritrova integralmente al terminale invertente:

$$v_1(t) = v_s(t)$$

Quindi la corrente che circola nel resistore R_1 è:

$$i_1(t) = \frac{v_s(t)}{R_1}$$

Mentre la corrente che circola nel resistore R_2 è:

$$i_2(t) = \frac{v_o(t) - v_s(t)}{R_2}$$

Pertanto:

$$\frac{v_s(t)}{R_1} = \frac{v_o(t) - v_s(t)}{R_2} \Rightarrow A_v = \frac{v_o(t)}{v_s(t)} = \frac{v_s(t) \left(1 + \frac{R_2}{R_1}\right)}{v_s(t)} = 1 + \frac{R_2}{R_1}$$

Lo stesso risultato può essere ottenuto considerando che, dal momento in cui c'è cortocircuito virtuale, **la corrente $i_i(t)$ è nulla e le resistenze R_1 e R_2 sono in serie**, costruendo un partitore di tensione per cui:

$$v_s(t) = v_o(t) \left(\frac{R_1}{R_1 + R_2}\right) \Rightarrow A_v = \frac{v_o(t)}{v_s(t)} = \frac{1}{\frac{R_1}{R_1 + R_2}} = 1 + \frac{R_2}{R_1}$$

Considerando due amplificatori ad operazionale, uno invertente e uno non invertente, caratterizzati dagli stessi parametri (resistenze uguali, amplificatori operazionali uguali e segnali uguali) non è vero che i guadagni dei due sono l'uno l'opposto dell'altro, come si potrebbe ingenuamente pensare; infatti:

$$1 + \frac{R_2}{R_1} \neq -\left(-\frac{R_2}{R_1}\right)$$

In genere, **in modulo amplifica di più un amplificatore non invertente**, sebbene quell'affermazione possa essere considerata valida per un rapporto tra resistenze abbastanza elevato.

Nella schematizzazione proposta per questo amplificatore **non è inclusa la resistenza serie al generatore di segnale perché il cortocircuito virtuale annulla qualsiasi caduta di tensione su di essa, rendendo la sua presenza ridondante e trascurabile.**

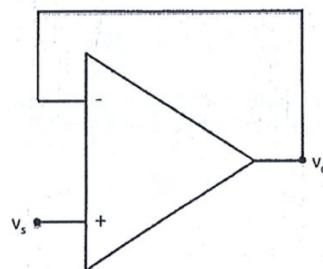
Lo scopo di questo amplificatore è di non invertire il segnale, a differenza di quando lo si amplifica in configurazione invertente ma, nonostante ciò, **non è stato analizzato un circuito perfettamente duale**; allora **non conveniva implementare due amplificatori ad operazionale** in configurazione invertente, il secondo dei quali con guadagno unitario in modulo, **per far elidere i segni negativi?** **Questa soluzione è possibile ma ben poco efficiente**, visto che nella configurazione non invertente la componente di guadagno costante (1) è nettamente inferiore all'usuale rapporto tra resistenze.

- **Buffer di tensione ad operazionale**

Per realizzare un buffer di tensione con un amplificatore operazionale è **necessario considerare quelle configurazioni per cui il guadagno sia unitario** (o prossimo all'unità) **a partire dalla configurazione non invertente** (che è molto più ideale, avendo resistenza di ingresso infinita grazie al cortocircuito virtuale):

$$A_v = 1 + \frac{R_2}{R_1} = 1 \Leftrightarrow \frac{R_2}{R_1} = 0 \Leftrightarrow R_2 = 0 \vee R_1 = \infty$$

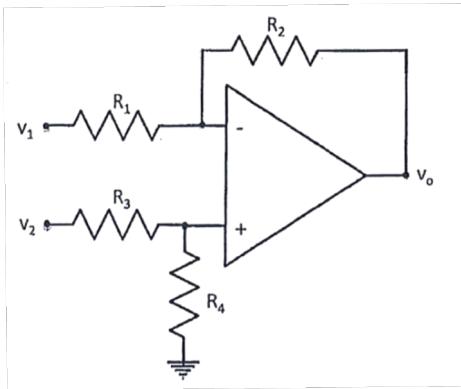
Quindi, **un amplificatore ad operazionale in configurazione non invertente si comporta come buffer di tensione se si sostituisce a R_1 un circuito aperto e a R_2 un cortocircuito**. Affinché il rapporto tra le due resistenze si annulli è sufficiente considerare anche solo numeratore nullo o solo denominatore infinito; tuttavia, poiché la resistenza implica comunque un dispendio di soldi, conviene implementare entrambe le soluzioni, risparmiando l'inserimento delle resistenze e realizzando il buffer di tensione come in figura:



- **Amplificatore differenziale ad operazionale**

Sfruttando le due configurazioni di amplificatore ad operazionale, invertente e non invertente, è possibile **realizzare un amplificatore differenziale ad operazionale**. Ma se è stato mostrato nei precedenti capitoli l'amplificatore differenziale, **perché è necessaria la sua versione ad operazionale?** Il motivo risiede in una **questione di stabilità: sui due terminali a cui è applicata la tensione, in un amplificatore operazionale deve sussistere una certa stabilità**, di cui l'amplificatore differenziale ad operazionale è meno sensibile.

Per analizzare il circuito **si preferisce usare la sovrapposizione degli effetti rispetto alle due tensioni in ingresso**, $v_1(t)$ e $v_2(t)$. Dette $v_{o1}(t)$ e $v_{o2}(t)$ le corrispondenti tensioni di uscita, l'amplificazione differenziale restituirà come risultato la tensione $v_o(t) = v_{o1}(t) + v_{o2}(t)$. Dal punto di vista grafico, **l'amplificatore differenziale si compone come segue**:



Si effettuino le seguenti prove:

$$\circ \quad v_1(t) \neq 0 \wedge v_2(t) = 0$$

Non circolando corrente nel terminale non invertente, non c'è caduta di tensione sulle resistenze R_3 e R_4 ; si riconosce, così, una semplice configurazione invertente:

$$v_{o1}(t) = -\frac{R_2}{R_1}v_1$$

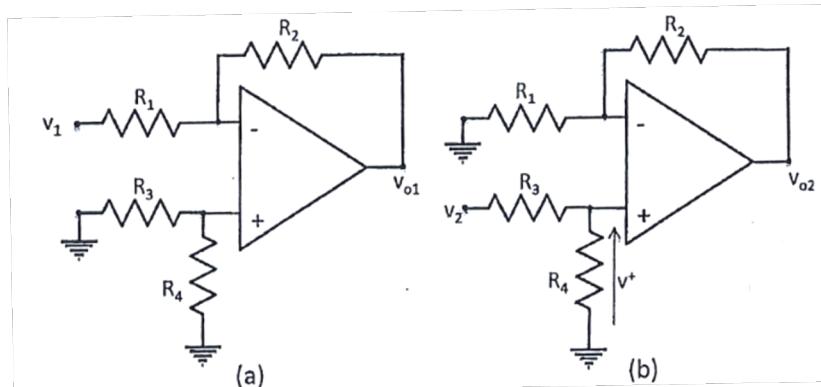
$$\circ \quad v_1(t) = 0 \wedge v_2(t) \neq 0$$

In questo caso, R_1 e R_2 formano la stessa rete di reazione; tuttavia, la tensione $v_2(t)$ non è applicata direttamente al morsetto non invertente, sul quale è applicata la tensione che insiste sulla resistenza R_4 (indicata con $v^+(t)$) ottenuta a partire dalla serie di R_3 e R_4 (visto che $i_i(t)$ è nulla):

$$v^+(t) = v_2(t) \cdot \frac{R_4}{R_3 + R_4} = v_2(t) \cdot \frac{R_4}{R_3} \frac{1}{1 + \frac{R_4}{R_3}}$$

Essendo in configurazione non invertente, l'uscita sarà:

$$v_{o2}(t) = v^+(t) \cdot \left(1 + \frac{R_2}{R_1}\right) = \frac{R_4}{R_3} \cdot \frac{1}{1 + \frac{R_4}{R_3}} \left(1 + \frac{R_2}{R_1}\right) v_2(t)$$



L'uscita totale del sistema sarà:

$$v_o(t) = v_{o1}(t) + v_{o2}(t) = -\frac{R_2}{R_1}v_1 + \frac{R_4}{R_3} \cdot \frac{1}{1 + \frac{R_4}{R_1}} \left(1 + \frac{R_2}{R_1}\right) v_2(t)$$

Che non è una differenza, bensì una combinazione lineare, finché non accade che:

$$\frac{R_4}{R_3} = \frac{R_2}{R_1}$$

In queste ipotesi:

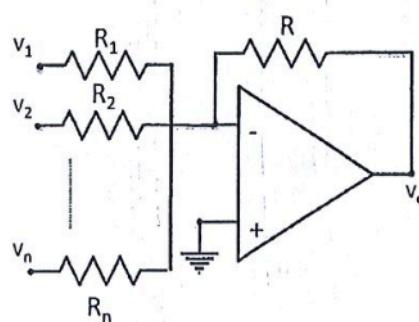
$$v_o(t) = \frac{R_2}{R_1} (v_2(t) - v_1(t))$$

Che è una differenza con guadagno differenziale A_d :

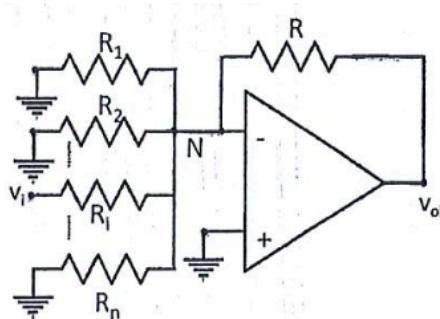
$$A_d = \frac{R_2}{R_1}$$

- **Sommatore ad operazionale**

Il sommatore ad operazionale è schematizzato come segue:



Per trovare la tensione di uscita quando sono presenti contemporaneamente gli n segnali da addizionare si ricorre nuovamente al principio di sovrapposizione degli effetti; per semplicità si considera solo la somma tra due segnali, da generalizzare ad n addendi. Quando solo l' i -esimo segnale è applicato in ingresso il circuito si configura come segue:



Grazie al cortocircuito virtuale, la tensione al nodo N è nulla, non scorre corrente nei resistori disaccoppiati e la configurazione è quella di un semplice amplificatore ad operazionale invertente:

$$v_{oi}(t) = -v_i(t) \frac{R}{R_i}$$

Risulta evidente come la tensione di uscita sia la somma di tutte le *n* componenti di questo circuito:

$$v_o(t) = -\left(\frac{R}{R_1} v_1(t) + \frac{R}{R_2} v_2(t) + \cdots + \frac{R}{R_n} v_n(t) \right)$$

Il sommatore realizzato è pesato dal rapporto tra la resistenza dell'amplificatore e le varie resistenze in serie ai generatori di segnale; se questi valori coincidono, il sommatore non è pesato da null'altro che dal segno –:

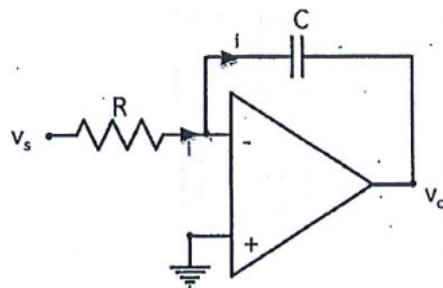
$$v_o(t) = v_1(t) + v_2(t) + \cdots + v_n(t)$$

In maniera del tutto analoga a quanto detto per l'amplificatore ad operazionale in configurazione non invertente, è possibile pensare ad un circuito più semplice in cui si realizza un sommatore utilizzando due amplificatori differenziali in serie o un differenziale con due sommatori in serie, senza introdurre nuove configurazioni; come nel caso precedente, questi dispositivi sono possibili ma ben poco efficienti, dal momento in cui è più economico ed elegante utilizzare i circuiti ad hoc piuttosto che una manipolazione di circuiti che non sono stati sviluppati per quello scopo.

- **Integratore ad operazionale**

Poiché le operazioni di derivata e di integrale sono due facce della stessa medaglia e poiché gli unici componenti di cui si ha memoria che sono modellabili tramite derivate sono gli induttori e i condensatori, si può intuire che integratore e derivatore ad operazionale debbano lavorare con suddetti dispositivi. In particolare, ciò che distingue i due circuiti sono le posizioni in cui vengono piazzati i condensatori (non si useranno molto gli induttori ma il principio di funzionamento è lo stesso).

Per un integratore il circuito è il seguente:



Si può dimostrare che, in questa configurazione, l'uscita è proporzionale all'integrale della tensione in ingresso; la corrente nel ramo R vale sempre:

$$i(t) = \frac{v_s(t)}{R}$$

Ma questa corrente circola nel condensatore, per cui vale la relazione:

$$i(t) = C \cdot \frac{dv(t)}{dt}$$

Ma la tensione ai capi del condensatore è proprio la tensione di uscita:

$$\frac{v_s(t)}{R} = C \frac{dv_o(t)}{dt}$$

Risolvendo questa equazione differenziale si ottiene la seguente soluzione:

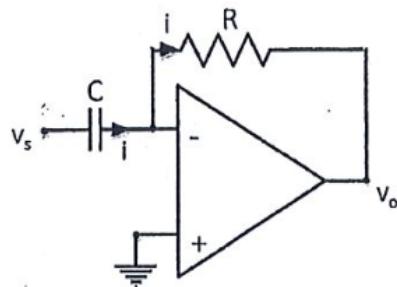
$$v_o(t) = -\frac{1}{RC} \int_0^t v_s(\tau) d\tau$$

Si noti che l'unità di misura della quantità sotto integrale è $V \cdot s$, che viene controbilanciata dalla quantità RC^{-1} misurata in s^{-1} .

Volendo essere pignoli, l'integrazione che viene fatta da questo circuito non è una vera e propria integrazione ma una derivazione inversa o, al più, sommatoria continua e lo suggeriscono gli estremi di integrazione, 0 e t ; infatti, il comportamento del condensatore è un comportamento cumulativo e l'integrale in esame può essere visto come una somma continua nel tempo.

- **Derivatore ad operazionale**

Per il derivatore ad operazionale, il condensatore è posto come in figura:



Ma le considerazioni a cui si giunge sono del tutto analoghe; infatti:

$$i(t) = C \frac{dv_s(t)}{dt}$$

$$v_o(t) = -Ri(t)$$

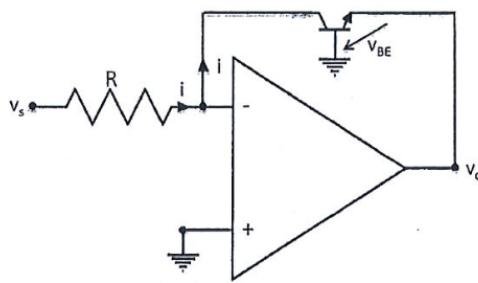
Per cui:

$$v_o(t) = -RC \frac{dv_s(t)}{dt}$$

Si possono fare le stesse considerazioni fatte per l'integratore riguardo l'unità di misura complessiva.

- **Amplificatore logaritmico ad operazionale**

L'ultima configurazione fondamentale che si prende in considerazione è l'**amplificatore logaritmico ad operazionale**; l'operazione di logaritmo è legata a quella di esponenziale e gli unici dispositivi in cui compare una caratteristica esponenziale sono il BJT e il diodo. Per come il circuito è composto, tuttavia, il BJT si comporta come diodo; infatti:



Si nota che i terminali di base ed emettitore sono allo stesso potenziale; quindi, la giunzione PN tra i due è modellabile tramite un circuito aperto, mentre quella tra Base e Collettore con un diodo. Grazie a questa equivalenza si preferisce, per una questione di semplicità, modellare il circuito come se al posto del BJT ci fosse un diodo.

La corrente che scorre nel diodo è la stessa che scorre nel resistore, quindi:

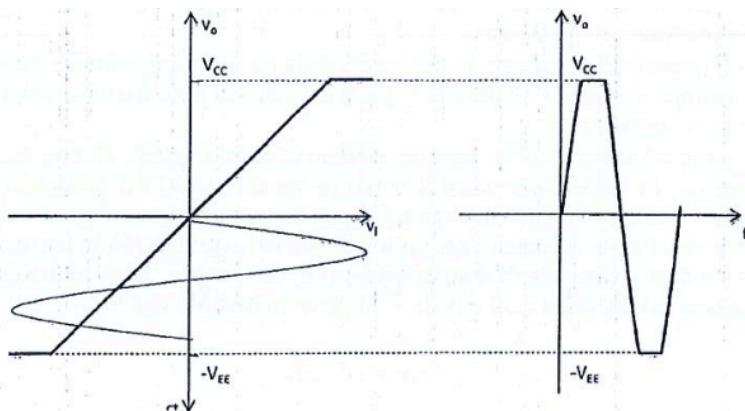
$$\frac{v_s(t)}{R} = I_o e^{-\frac{v_o(t)}{V_t}}$$

Da cui:

$$v_o(t) = -V_t \log\left(\frac{v_s(t)}{RI_0}\right)$$

Da questa relazione si può osservare una imprevedibile ed enorme sensibilità del dispositivo nei confronti delle variazioni di $v_s(t)$ a causa della presenza di I_0 al denominatore dell'argomento del logaritmo; se si volesse controbilanciare questo effetto, sarebbe necessaria una resistenza dell'ordine delle migliaia di Giga ohm ($I_0 \approx 10^{-12} A$), valori praticamente irraggiungibili. Questo fenomeno non è eliminato quando al posto di un diodo si modella il circuito con un BJT, solo che non è altrettanto evidente analiticamente. Segue che l'amplificatore logaritmico ad operazionale non è la soluzione migliore se si vuole restituire il logaritmo di un segnale.

Le ipotesi di linearità su cui è stata costruita tutta la trattazione dell'amplificatore operazionale non sono state ancora enunciate, è stato solo mostrato il modello equivalente dell'operazionale senza soffermarsi su quando tale approssimazione fosse possibile. Il limite sull'ampiezza dei segnali che è possibile applicare in ingresso all'operazionale è imposto dall'ovvia considerazione che la tensione di uscita non può in alcun modo superare il valore delle tensioni di alimentazione; di conseguenza, la caratteristica di trasferimento ingresso – uscita di un circuito operazionale assume la forma mostrata nella figura che segue:



Imponendo che:

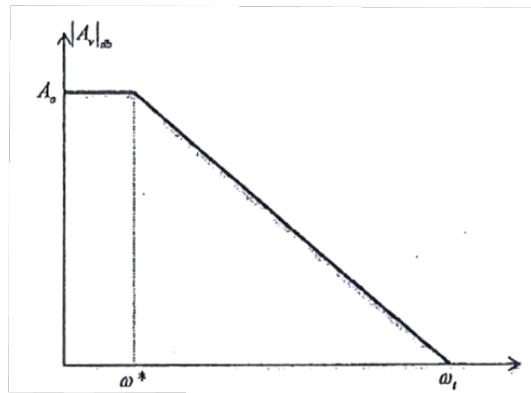
$$v_o(t) \in [-V_{EE}, V_{CC}]$$

Nel caso in cui la tensione di uscita sforasse teoricamente (superiormente o inferiormente) questi valori, essa sarebbe praticamente limitata all'estremo sforato dell'intervallo. La condizione di linearità appena enunciata prende anche il nome di **condizione di saturazione dell'amplificatore operazionale**.

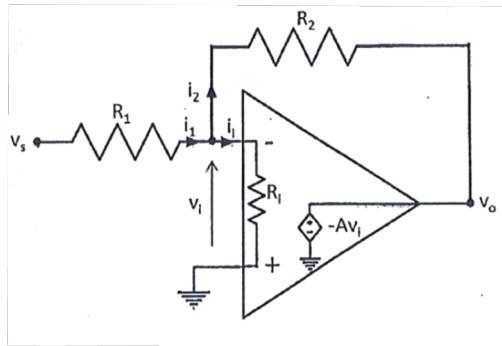
COMPORTAMENTO IN FREQUENZA DELL'OPERAZIONALE

Tra le cinque ipotesi di linearità enunciate ed usate in precedenza, quella **meno vicina alla realtà è l'ipotesi di banda passante infinita**; in realtà è anche **poco desiderabile**, visto che bande passanti infinite amplificano anche i rumori esterni. Più in generale, la **combinazione di elevato guadagno ad elevata frequenza è una condizione intrinsecamente instabile**, cioè che porta la tensione di uscita dell'amplificatore a divergere indipendentemente dal segnale applicato in ingresso; questa instabilità, in realtà, è **presente anche a basse frequenze**, visto che il **guadagno dell'operazionale è estremamente alto**.

Studiando l'amplificatore, **non ci si è soffermati sulla determinazione della frequenza di taglio superiore della banda**, limitandosi a dire che essa è infinita; il motivo è che, **nella pratica**, viene aggiunto un **capacitore di capacità abbastanza elevata** (in realtà, la capacità è modesta ma ponendola sul terzo stadio viene amplificata sul quarto) **da dominare la risposta dell'amplificatore alle alte frequenze** (diventando un cortocircuito). Da queste operazioni si ottengono **due risultati: in primis si rende perfettamente predicibile la frequenza di taglio superiore**, svincolandola dalle capacità interne parassite dei transistori, in secondo luogo **essa viene fissata ad un valore sufficientemente basso da evitare i problemi di instabilità precedentemente menzionati**. Con queste scelte, **la risposta in frequenza dell'operazionale a ciclo aperto assume la seguente forma**:



Nei dispositivi commerciali, ω^* è intorno alla decina di Hertz e si può ragionevolmente pensare che siano limitati all'elaborazione di segnali continui o quasi. Per convincersi del contrario, si esamini la configurazione invertente dell'amplificatore ad operazionale studiata in precedenza:



Si valuti la risposta in frequenza di questo amplificatore nel caso in cui l'operazionale sia caratterizzato dalla risposta a ciclo aperto in questione. La forma analitica che genera tale andamento è la seguente:

$$A(i\omega) = \frac{A_o}{1 + i \frac{\omega}{\omega^*}}$$

Per valutare le prestazioni dell'amplificatore a ciclo chiuso è, chiaramente, inutile considerare le approssimazioni dell'amplificatore ideale, dal momento in cui si vuole valutare proprio l'effetto di un guadagno non infinito che varia con la frequenza.

Sostituendo questa relazione alla formula di determinazione del guadagno dell'amplificatore ad operazionale:

$$A_v(i\omega) = \frac{v_o(t)}{v_s(t)} = -\frac{R_2}{R_1} \cdot \frac{1}{1 + \frac{1}{A_o} \left(1 + i \frac{\omega}{\omega^*}\right) \left(1 + \frac{R_2}{R_1}\right)}$$

Trascurando il termine R_2/R_1 per semplicità. Ponendo $A_v(0) = -R_2/R_1$ e trascurando l'unità rispetto a R_2/R_1 , si può semplificare questa formula in:

$$A_v(i\omega) \approx \frac{A_v(0)}{1 + \frac{1}{A_o} \left(1 + i \frac{\omega}{\omega^*}\right) \frac{R_2}{R_1}} \approx \frac{A_v(0)}{1 + \frac{1}{A_o} \frac{R_2}{R_1} + i \frac{1}{A_o} \frac{\omega}{\omega^*} \frac{R_2}{R_1}} \approx \frac{A_v(0)}{1 + i \frac{1}{A_o} \frac{\omega}{\omega^*} \frac{R_2}{R_1}} = \frac{A_v(0)}{1 + i \frac{\omega}{\omega_H}}$$

Si è ricavato che la forma della risposta in frequenza dell'amplificatore invertente ad operazionale è ancora a singolo polo, con la frequenza di taglio superiore pari a:

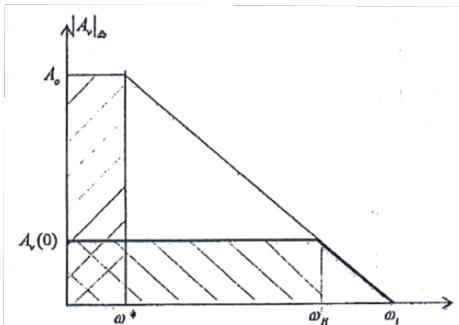
$$\omega_H = \omega^* \frac{A_o}{A_v(0)}$$

Con questa formula si può dire che **la frequenza di taglio del sistema reazionato è data dalla frequenza di taglio superiore del sistema non reazionato moltiplicata per il rapporto tra il guadagno in continua a ciclo aperto ed il guadagno in continua a ciclo chiuso.** Dal momento in cui il rapporto dei guadagni è, generalmente, molto elevato, si può concludere che **la frequenza di taglio superiore del sistema reazionato è molto più grande di quella del sistema non reazionato.**

La formula appena enunciata può essere alternativamente **messa sotto forma di prodotto di guadagno per banda costante:**

$$\omega_H \cdot A_v(0) = \omega^* \cdot A_o$$

Con il quale è possibile dire che, **aumentando il guadagno rispetto alla configurazione a ciclo aperto, di tanto aumenta la banda passante della configurazione a ciclo chiuso**. Graficamente, la risposta in frequenza delle due configurazioni assume la forma seguente:



Si noti che le due aree evidenziate sono uguali, proprio come apprezzato analiticamente dalla formula di prodotto di guadagno costante. Tuttavia, le formule non permettono di notare che la pulsazione ω_t , che rappresenta il punto in cui il modulo del guadagno diviene unitario (la scala è logaritmica, il guadagno unitario corrisponde all'intersezione con l'asse delle ascisse), è anch'essa un'invariante del sistema proprio come l'area sottesa alla banda. Questa pulsazione, detta frequenza di transizione, è proprio il valore del prodotto guadagno per banda; infatti:

$$A(i\omega) = \frac{A_o}{1 + i \frac{\omega}{\omega^*}} = 1 \Leftrightarrow \omega_t = \omega^* \cdot A_o$$

$$A_v(i\omega_t) = \frac{A_v(0)}{1 + i \frac{\omega_t}{\omega_H}} = 1 \Leftrightarrow \omega_t = \omega_H \cdot A_v(0)$$

Si ricordi che pulsazione e frequenza, sebbene riferiscano concetti simili e proporzionali, sono due grandezze differenti, legate dalla relazione:

$$\omega = 2\pi f$$

LA NON IDEALITÀ DELL'AMPLIFICATORE OPERAZIONALE

Nei circuiti ad operazionale possono occorrere dei fenomeni poco graditi che sono causati da condizioni di non idealità dell'amplificatore ad operazionale o da precise condizioni di funzionamento (questi ultimi sono più prevedibili). Di seguito saranno analizzati alcuni di questi fenomeni nel dettaglio e, lì dove possibile, saranno anche proposte delle soluzioni.

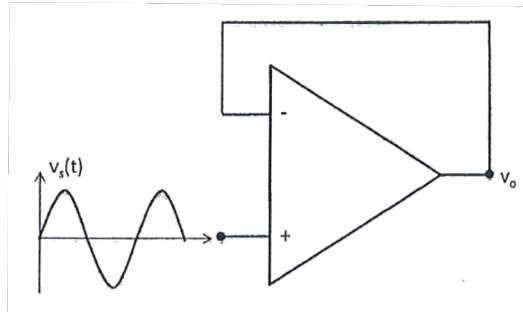
SLEW RATE

In precedenza, è stato detto che la risposta in frequenza dell'operazionale è dominata dalla capacità elevata inserita fisicamente e volontariamente nel circuito; tale capacità ha il compito di abbattere il guadagno dell'amplificatore alle alte frequenze e, pertanto, la sua posizione nel circuito è necessariamente in parallelo al percorso del segnale (in ingresso o in uscita) in modo da cortocircuitarlo verso massa quando la frequenza cresce.

Questa capacità, oltre che a determinare la frequenza di taglio superiore, **pone anche un limite alla massima velocità di variazione della tensione di uscita** (ovvero alla corrente che può circolare):

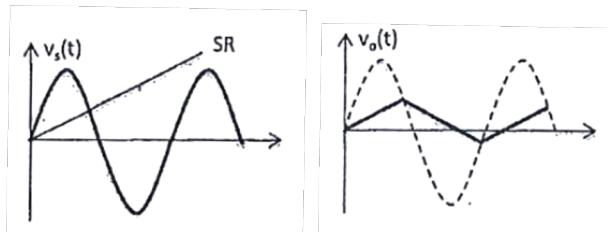
$$\left. \frac{dv(t)}{dt} \right|_{max} = \frac{i_{max}(t)}{C}$$

Questo valore di massima variazione dell'uscita è definito **slew rate (SR)**. Per studiare agevolmente il fenomeno, si consideri il seguente amplificatore ad operazionale in configurazione di buffer di tensione, al cui ingresso è applicato un segnale sinusoidale:

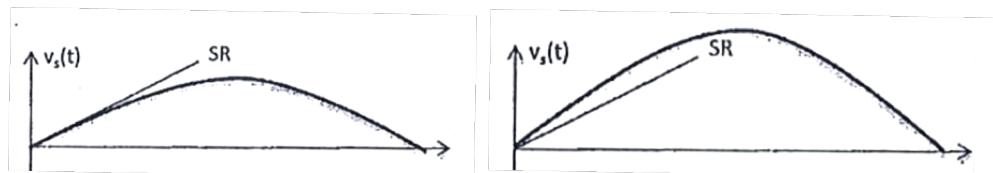


Per definizione, questo amplificatore ripropone in uscita il segnale in ingresso identicamente; quindi $v_o(t) = v_s(t) \forall t$. Si supponga, però, che l'operazionale sia caratterizzato da una slew rate rappresentata da una retta nel piano $v(t) - t$.

Nel caso in cui la pendenza della slew rate sia inferiore alla variazione del segnale di ingresso (e quindi di uscita), il segnale in uscita non potrà essere la perfetta riproduzione di quello di ingresso; l'amplificatore non riesce ad inseguire il segnale di ingresso e l'uscita diventa un'onda triangolare:



Questo effetto è legato contemporaneamente sia all'ampiezza del segnale che alla sua frequenza; infatti, di seguito sono riportati due segnali di ampiezza uguale ma frequenza diversa: il primo ha una frequenza tale da, con quell'ampiezza, non superare la slew rate, mentre il secondo è più frequente e riesce a superare la slew rate. Il primo segnale non è distorto in uscita, il secondo sì.



TENSIONE DI OFFSET

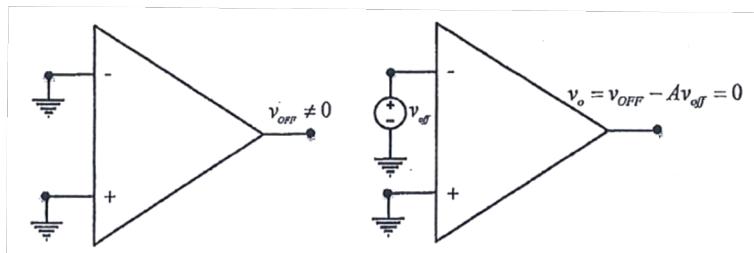
La relazione costitutiva di un operazionale è la seguente:

$$v_o(t) = A(v_2(t) - v_1(t))$$

Supponendo che, quando $v_2(t) = v_1(t)$, la tensione di uscita è nulla; quindi, si presuppone che quando ai due terminali di ingresso (invertente e non invertente) è applicato lo stesso segnale, in uscita non può essere rilevata alcuna tensione. La realtà è ben diversa e può svilupparsi, in queste condizioni, un'uscita non identicamente nulla; la tensione in questione è detta **tensione di offset in uscita** e deriva, in primo luogo, dalla struttura interna di un operazionale.

Qualora gli stadi differenziali che compongono l'operazionale non siano perfettamente bilanciati (ovvero i parametri costitutivi dei transistori e le resistenze interne non siano perfettamente uguali), **può svilupparsi tra i collettori dei transistori una differenza di potenziale diversa da zero, dovuta al semplice passaggio della corrente statica di polarizzazione** (che produce effetti diversi nei due rami del differenziale quando questi non sono simmetrici).

Il fenomeno della tensione di offset è inevitabile, dal momento in cui anche in elettronica integrata è estremamente improbabile la creazione di due dispositivi perfettamente identici (si sta lavorando a livello atomico, è molto difficile direzionare un atomo) e, quindi, **in tutti gli operazionali la tensione in corrispondenza di ingressi nulli è, seppur molto piccola, non nulla**. Nei dispositivi commerciali è solito trovare dei **terminali aggiuntivi che hanno lo scopo di correggere la tensione di offset**, applicando in ingresso una tensione di valore opposto a quella di offset determinata, detta **tensione di offset in ingresso**; tuttavia, poiché il segno dell'offset in uscita non è noto a priori, questa correzione non può che essere fatta empiricamente.



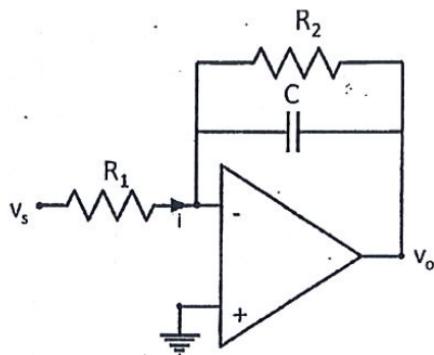
Anche se la stessa modifica può essere fatta sull'uscita con un generatore di tensione in serie che eroga la stessa tensione di quella rilevata di offset.

È anche possibile, però, che sia la topologia stessa del circuito a generare una tensione di uscita non nulla pur con ingressi nulli.

L'INTEGRATORE REALE E LA RETE DI REAZIONE GENERALIZZATA

Il circuito integratore precedentemente mostrato soffre di problemi di instabilità che lo rendono, di fatto, inutilizzabile e non funzionante. Poiché negli operazionali reali è inevitabile la presenza di una tensione continua di offset, il segnale in ingresso è traslato verso l'alto (o il basso) e si può spesso trovare in condizioni di saturazione, restituendo così in uscita sempre V_{CC} (o $-V_{EE}$).

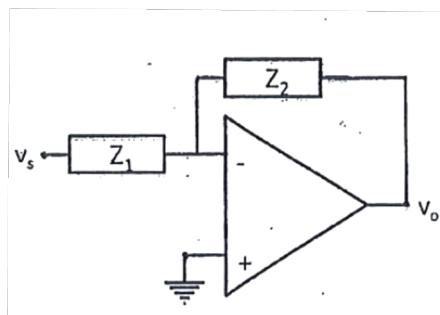
Un integratore reale, che risolve questi problemi di stabilità, si compone come segue:



Le sovrapposizioni della tensione di offset sono, così, limitate: il condensatore si comporta come circuito aperto per il segnale costante, il quale è forzato a seguire la strada dell'amplificazione (a basso guadagno) mentre il segnale di ingresso quella dell'integrazione.

La rete di reazione dell'amplificatore operazionale non deve essere necessariamente resistiva (come visto per l'integratore e il derivatore ad operazionale). Essa può, infatti, essere costituita da generiche impedenze, ampliando notevolmente il numero di applicazioni possibili.

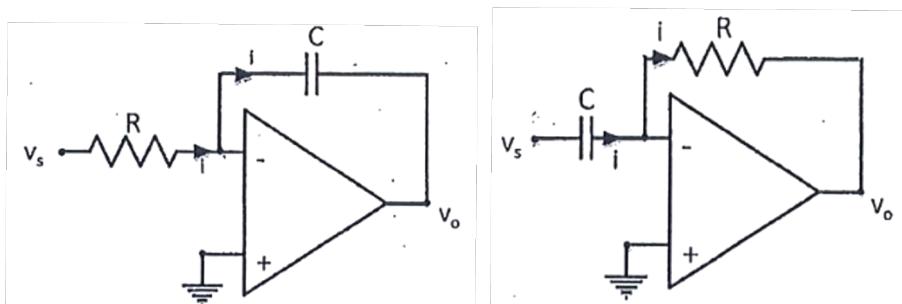
Per questo tipo di circuiti **non conviene l'analisi nel dominio del tempo, è più efficace quella nel dominio di Laplace**; come esempio, si consideri la configurazione invertente con due impedenze Z_1 e Z_2 al posto di R_1 e R_2 :



La funzione di trasferimento che esprime il guadagno di tensione risulta essere:

$$A_v = -\frac{Z_2}{Z_1}$$

Per rendere concreto quanto si sta per dire, si può particolarizzare il circuito in figura nei due seguenti:



Nel primo caso:

$$A_v = -\frac{\frac{1}{sC}}{R} = -\frac{1}{RCs}$$

Mentre nel secondo caso:

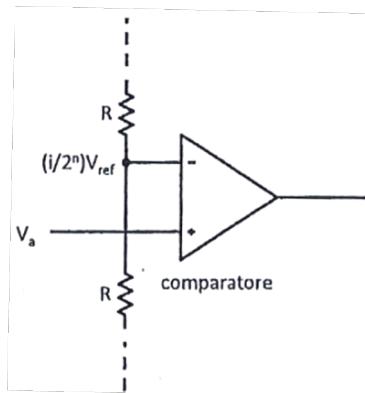
$$A_v = -\frac{\frac{R}{1}}{\frac{1}{sC}} = -RCs$$

Ricordando che **nel dominio di Laplace, la divisione per s corrisponde all'integrazione e la moltiplicazione per s alla derivazione**; infatti, i due circuiti corrispondono, rispettivamente, all'integratore e al derivatore precedentemente menzionati.

ELETTRONICA DIGITALE

IL CONVERTITORE FLASH

Per convertitore **flash** (o in parallelo) a n bit si intende un circuito composto da 2^{n-1} comparatori ed un codificatore. Un segnale V_a da convertire viene applicato agli ingressi non invertenti dei comparatori, mentre quello invertente è connesso ad una rete che ripartisce la tensione di riferimento in 2^n fasce in modo da fissare i livelli di riferimento; ciascun comparatore poi commuta la sua uscita a 1 quando V_a supera il rispettivo livello di riferimento. Quindi, a valle dei 2^{n-1} comparatori, è possibile inserire una opportuna rete di codifica $2^n - n$ che rappresenta la parola di n bit corrispondente al dato acquisito. Lo schema del generico comparatore, corrispondente ad uno dei 2^{n-1} stadi, è così realizzato:



Questo tipo di convertitore **consente un'elevata velocità di conversione** (dipende dalla somma dei tempi di propagazione del comparatore e della rete di codifica); **gli unici inconvenienti di questo tipo di convertitore sono l'elevato numero di comparatori richiesti per alte risoluzioni** e la necessità di realizzare una rete resistiva con 2^n resistenze uguali tra loro, ponendo evidenti **problemi di accuratezza**.

INVERTITORI LOGICI

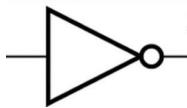
Dal punto di vista elettronico, **un circuito è detto digitale quando i segnali di ingresso, di uscita e intermedi possono assumere solo due valori distinti**, che rappresentano il corrispettivo booleano di 0 e 1; particolare attenzione va posta a tutti i valori intermedi che tali segnali possono assumere, visto che **la loro variazione è continua e non discreta come quella logica tra 0 e 1**.

In algebra booleana è definito **invertitore logico** una qualsiasi porta che riceve in ingresso un **informazione binaria** e ne restituisce il complementare; quindi, se in ingresso è posto 0 in uscita sarà restituito 1 e viceversa. **Il ruolo dell'invertitore è fondamentale**, non solo dal punto di vista tecnico ma anche realizzativo: verrà mostrato a breve come **anche i circuiti digitali più complessi saranno realizzati a partire da un invertitore logico elementare, le cui prestazioni andranno a condizionare il funzionamento di tutto il circuito**; quindi, sia l'analisi che la sintesi di un circuito digitale qualsiasi, più o meno complesso, parte dall'invertitore logico elementare.

Algebricamente, una tale porta logica è indicata con la seguente notazione:

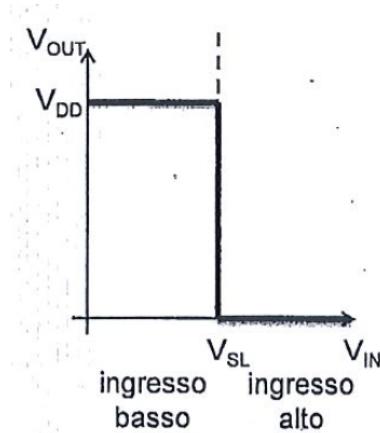
$$Y = \neg A = \bar{A}$$

A	Y
0	1
1	0



Tuttavia, **questo livello di astrazione non fornisce alcun suggerimento sulla possibile realizzazione circuitale del corrispettivo dispositivo elettronico.** Per approcciare produttivamente al problema è necessario indagare il significato elettrico dei valori booleani 0 e 1; in particolare, essi non saranno assegnati a 0V e 1V (si parlerà tendenzialmente di tensioni, meno di correnti) ma alla minima e alla massima tensione di esercizio del sistema: lo 0 logico corrisponderà alla tensione $V = 0$ e l'1 logico alla tensione di alimentazione $V = V_{DD}$. Si può, quindi, dire che un invertitore digitale prende in ingresso una tensione nulla (o una tensione di alimentazione) e restituisce la tensione di alimentazione (o una tensione nulla).

In realtà, la definizione di un invertitore digitale è più complessa e prende in considerazione alcune proprietà di incredibile potenza dei circuiti elettronici digitali; infatti, è possibile definire un **livello di tensione massimo entro cui l'uscita permane nel proprio stato logico**, la tensione di soglia V_{SL} . Il circuito che restituisce una tensione nulla per tutti i valori di tensione di ingresso inferiori alla tensione di soglia e la tensione di alimentazione per tutti i valori di tensione di ingresso superiori alla tensione di soglia è detto **invertitore logico ideale**. La **caratteristica di trasferimento** (o di ingresso – uscita) permette di visualizzare graficamente questa definizione; è possibile individuare nel grafico una tripartizione: la prima e la terza zona sono quelle in cui la tensione di ingresso è, rispettivamente, minore e maggiore (ma non uguale) alla tensione di soglia e, in corrispondenza di queste regioni, l'uscita inverte l'ingresso, mentre la variazione dell'uscita è nulla. La zona di transizione corrispondente a tensioni pari a quella di soglia presenta, per questa configurazione ideale, **una pendenza perfettamente verticale** (evidenziando i limiti di questo modello) e corrisponde ad una **commutazione immediata dell'uscita**.

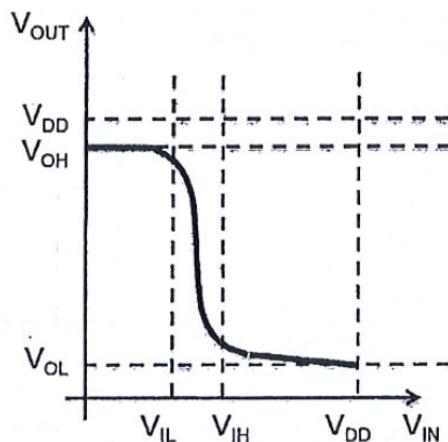


Con queste informazioni a disposizione è possibile già individuare la **peculiarità dei dispositivi elettronici digitali**: essi producono, in un intervallo di valori di ingresso, un'uscita indipendente dalle variazioni dell'ingresso; ovvero, il dispositivo si rivela essere insensibile a variazioni del segnale di ingresso inferiori (al più) a $\Delta V = V_{SL}$. Ciò significa che questi dispositivi non vengono turbati minimamente da leggeri rumori o piccole interferenze esterne; nel corso della trattazione verranno ignorati questi inconvenienti perché, essendo particolarmente resilienti e composti da

diversi stadi in serie, i circuiti digitali più articolati li elimineranno automaticamente al primo stadio e non saranno propagati ai successivi.

A rigore, in un circuito digitale non è detto che la massima tensione di uscita coincida con la tensione di alimentazione, può capitare che sia inferiore (mai superiore, non esiste energia gratis); analogamente, la minima tensione di uscita può non essere quella nulla, può anche essere $-V_{DD}$. Anche in questo caso reale, però, è possibile definire le soglie inferiori e superiori che rappresentano i livelli logici del circuito; si denota con V_{OH} (Output High) come il livello logico nominale alto e con V_{OL} (Output Low) il livello logico nominale basso. Con queste definizioni in considerazione si è superati la prima idealità dell'invertitore logico ideale, ovvero quella relativa alla corrispondenza tra tensioni di uscita e valori booleani; resta da "realizzare" la pendenza infinita a cavallo della tensione di soglia.

Il tratto a pendenza verticale nella realtà pratica è una soluzione quasi irrealizzabile (o almeno eccessivamente costosa) e, pertanto, totalmente ideale; il modo più accurato di modellare una situazione di questo tipo può consistere in un tratto a pendenza estremamente elevata (ma che non sarà mai verticale):



Questa curva, sebbene possa essere il più ripida possibile, va a perdere la proprietà filtranti dell'invertitore ideale per cui fino alla tensione di soglia era mantenuto un determinato valore di uscita (il circuito filtra di meno i rumori); bisogna, quindi, determinare il valore di tensione per cui l'uscita di un invertitore può essere considerata alta anche se diversa da V_{OH} o bassa anche se diversa da V_{OL} .

Si consideri la pendenza della caratteristica di trasferimento come il guadagno di un amplificatore analogico e la si divida in tre regioni:

1. Tra 0 e V_{IL} (Input Low), in cui la pendenza è in modulo minore di uno;
2. Tra V_{IL} e V_{IH} (Input High), in cui la pendenza è in modulo maggiore dell'unità;
 - a. In questo intorno le capacità filtranti della porta logica sono piuttosto scarse e, pertanto, va funzionalmente evitato e ridotto al minimo;
3. Tra V_{IH} e V_{DD} , in cui la pendenza è in modulo di nuovo minore di uno.

Quindi, si individuano in V_{IL} e in V_{IH} i punti in cui la pendenza della caratteristica (in termini più formali si potrebbe parlare di derivata) è pari a -1. Questa definizione, in combinazione con quella di valori logici di tensione nominale, fornisce uno strumento più che sufficiente per definire il concetto di margine di rumore (Noise Margin):

$$NM_L = V_{IL} - V_{OL} \wedge NM_H = V_{OH} - V_{IH}$$

Essi **quantificano la capacità di un circuito digitale di respingere un'eventuale perturbazione sovrapposta all'ingresso**, indicando **quanto siano ampie le regioni di confidenza dell'ingresso di un circuito digitale** (quanto ampio possa essere il livello di rumore senza che i segnali di ingresso e di uscita vengano ad essere alterati). Tranne in alcuni casi, **il margine di rumore superiore è diverso da quello inferiore ed il più piccolo dei due condiziona inevitabilmente tutto il funzionamento della porta logica**; infatti, si definisce come **ampiezza massima di disturbo** come il più piccolo tra i due margini di rumore:

$$NM = \min\{NM_L, NM_H\}$$

Si possono già distinguere le **differenze principali tra un invertitore ideale ed uno reale**:

Invertitore ideale	Invertitore reale
$V_{IL} = V_{IH} = V_{SL}$	$V_{IL} < V_{IH}$
$NM_L = NM_H = NM = V_{SL}$	$NM_L \neq NM_H$

Al giorno d'oggi **uno dei fattori che più limita la performance di un circuito digitale è la dissipazione della potenza** e, pertanto, è necessario fare alcune digressioni sull'argomento. La potenza dissipata è definita come:

$$P_{diss} = \frac{1}{T} \int_T^{\square} V_{DD} \cdot i(t) dt$$

Con **$i(t)$ la corrente assorbita dalla porta logica durante il suo funzionamento e T un intervallo di funzionamento che comprende la permanenza della porta al livello logico basso e alto**. Nell'equazione, ci si riferisce al valore medio nel tempo della potenza dissipata, in quanto è questa che viene trasformata in calore per effetto Joule e che deve essere, pertanto, dissipata da opportuni sistemi di smaltimento del calore. È uso comune **distinguere i contributi relativi alla potenza dissipata staticamente dal circuito nei suoi due stati stabili e quella necessaria nelle commutazioni tra questi ultimi**:

$$P_{diss} = P_{DC} + P_{AC}$$

La **potenza dissipata statica**, a sua volta, la si può esprimere **valutando separatamente le potenze dissipate quando la porta logica si trova nello stato alto o in quello basso** attraverso la loro media:

$$P_{DC} = \frac{P_H + P_L}{2} = \frac{V_{DD}(I_H + I_L)}{2}$$

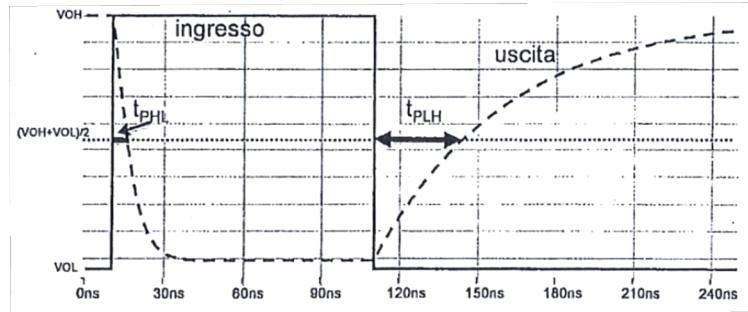
Sulla potenza dissipata durante le fasi transitorie non ci si sofferma più di tanto perché quando $P_{DC} \neq 0$ si può ragionevolmente pensare che il suo contributo sia preponderante e approssimare con buon margine di errore la potenza dissipata a:

$$P_{diss} \approx P_{DC}$$

Ovviamente, quando in ingresso avviene la commutazione di un segnale digitale, l'uscita non cambierà istantaneamente valore per adattarsi al nuovo ingresso, bensì richiederà un determinato tempo di adattamento. Per valutare la capacità di un circuito digitale di rispondere a segnali variabili

rapidamente non si utilizzano i concetti di risposta in frequenza o di tempo di salita/discesa tipici della teoria dei sistemi lineari o dell'elettronica analogica, si preferiscono strumenti meno legati alle caratteristiche interne del circuito ma facilmente misurabili e quantificabili analizzando le forme d'onda relative ai segnali di ingresso e uscita del circuito.

Riferendosi sempre ad un invertitore, si analizzi la forma d'onda presentata in figura, tenendo in considerazione che la linea piena rappresenta il segnale di ingresso (fatto quanto più ideale possibile per una questione di semplicità) e la linea tratteggiata la relativa uscita:



Posizionata una soglia al valore medio tra V_{OL} e V_{OH} , si definisce **tempo di propagazione alto – basso t_{PHL}** l'intervallo di tempo necessario affinché l'uscita passi per la soglia dopo una transizione basso – alto dell'ingresso; dualmente, il **tempo di propagazione basso – alto t_{PLH}** è definito come l'intervallo di tempo necessario affinché l'uscita passi per la soglia dopo una transizione alto – basso. Genericamente, si parla di **tempo di propagazione** (o ritardo di propagazione) di un generico circuito logico facendo riferimento al **valor medio di queste due quantità**:

$$t_p = \frac{t_{PLH} - t_{PHL}}{2}$$

È importante sottolineare che, proprio come il **margine di rumore**, anche i tempi di propagazione non sono uguali se non in casi particolari (non strettamente connessi) e che il più alto dei due andrà a limitare il comportamento dinamico del circuito; inoltre, per circuiti digitali in cascata, il **tempo di propagazione ha la proprietà**, particolarmente fastidiosa, di essere additivo.

Il tempo di propagazione è, probabilmente, la **grandezza dinamica più importante per i circuiti logici**, dal momento in cui **delimita inferiormente la massima durata di un segnale applicabile ad un circuito** (se il segnale in ingresso permane al valore logico alto o basso per un tempo inferiore a t_p il circuito non ha il tempo necessario a commutare l'uscita e, quindi, risulta insensibile al segnale di ingresso stesso); in altre parole, il **tempo di propagazione descrive la massima frequenza con cui possono variare i bit in ingresso ad una porta logica**.

Una valutazione del costo di un circuito digitale veloce in relazione alla potenza dissipata può essere fornita dal **prodotto ritardo – potenza**:

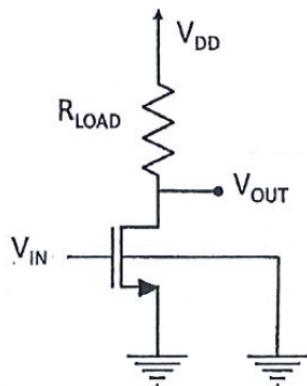
$$PD = P_{diss} \cdot t_p$$

Ogni porta logica possiede dei limiti riguardo al numero di ingressi che può prevedere (fan – in) e al numero di porte che può pilotare in uscita (fan – out); tali limiti sono imposti con il fine di mantenere la degradazione del segnale in uscita o le prestazioni della porta logica a valori accettabili. I limiti in questione sono da determinare sul singolo caso, generalmente si può solo dire che **all'aumentare delle porte logiche collegate in uscita aumenta inevitabilmente il tempo di**

propagazione (aumenta la capacità di carico complessiva) e che all'aumentare del numero di ingressi possono degradarsi i margini di rumore perché può aumentare, ad esempio, la V_{OL} .

Nell'ambito dell'elettronica digitale **i dispositivi che fungono da interruttori**, particolarmente utili quando si parla di invertitori, **sono i MOSFET**: con riferimento alla tensione di Gate come **parametro di controllo**, questo dispositivo si comporta da **circuito aperto** quando è interdetto (la corrente è virtualmente nulla) e come **cortocircuito** nella parte lineare della regione di triodo; pertanto, **pilotando la Gate opportunamente, si può replicare il comportamento di un interruttore tramite un MOSFET**.

Un primo **invertitore reale** è quello **a carico resistivo**, riassunto in figura:



In cui si è voluto esplicitare il **body a massa**. In fase di progettazione della porta logica, **la tensione di alimentazione V_{DD} e i drogaggi del MOSFET sono a carico della particolare tecnologia che si deve utilizzare**, mentre **il progettista ha un margine solo sulla resistenza di carico R_{LOAD} e sulle dimensioni geometriche del MOSFET, W e L** .

Volendo valutare, dapprima analiticamente, il comportamento di un circuito di questo tipo, si ricordi che **la caratteristica di trasferimento lega dal punto di vista statico le tensioni di uscita ad ogni possibile valore della tensione di ingresso**; pertanto, facendo variare V_{IN} dal valore più basso (0V) a quello più alto (V_{DD}), si può ottenere un risultato soddisfacente. Si può, in primis, considerare che, finché la tensione V_{IN} non supera la tensione di soglia del MOSFET, non può circolare alcuna corrente nel dispositivo, la caduta sul carico sarà nulla e l'uscita sarà trasparente a tutto il dispositivo:

$$V_{OUT} = V_{DD} - R_{LOAD} \cdot 0 = V_{DD}$$

Ottenendo un risultato piuttosto utile:

$$V_{OH} = V_{DD}$$

Appena si pone in ingresso una tensione $V_{IN} > V_{th}$, il MOSFET entra in conduzione e si troverà in **pinch-off** ($V_{DS} > V_{GS} - V_{th}$); in questa regione, **il legame tra corrente di Drain e tensione di ingresso V_{IN} è quadratico**:

$$I_D = \frac{1}{2} \mu_n C_{ox} \cdot \frac{W}{L} (V_{IN} - V_{th})^2$$

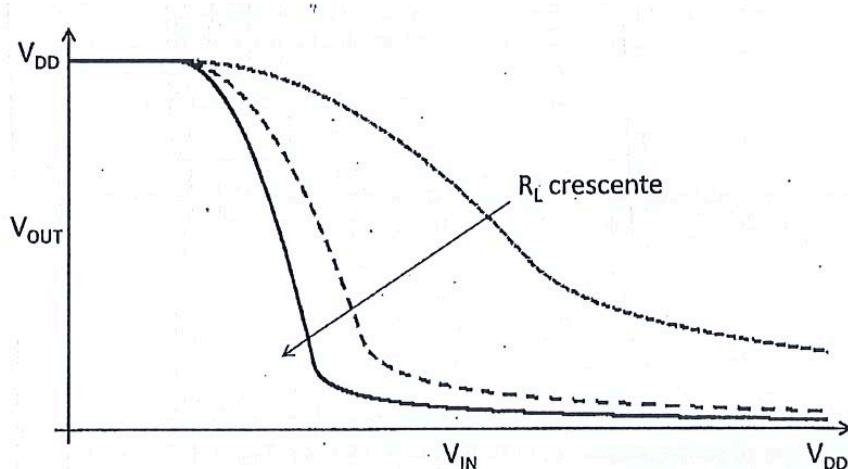
Pertanto, la **tensione di uscita**:

$$V_{OUT} = V_{DD} - \frac{1}{2} R_{LOAD} \mu_n C_{ox} \cdot \frac{W}{L} (V_{IN} - V_{th})^2$$

Che è una **parabola con concavità rivolta verso il basso**. Quando, poi, l'ingresso sarà sufficientemente elevato per portare il MOSFET in triodo:

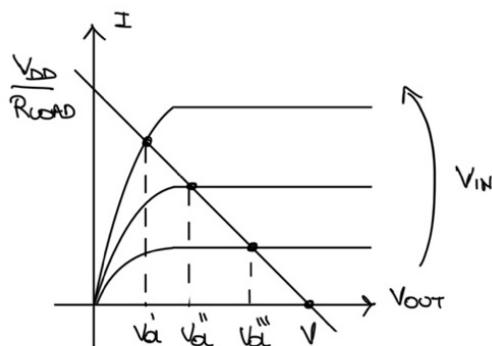
$$V_{OUT} = V_{DD} - \frac{1}{2} R_{LOAD} \mu_n C_{ox} \cdot \frac{W}{L} [2(V_{IN} - V_{th})V_{OUT} - V_{OUT}^2]$$

Mettendo insieme i risultati ottenuti:



Si noti che, all'aumentare di R_{LOAD} , diminuisce la tensione V_{OL} e aumentano i margini di rumore.

La valutazione analitica di V_{OL} è leggermente più articolata e deve tenere in considerazione il fatto che il suo valore non deve portare eventuali stadi successivi collegati in cascata in conduzione; pertanto, si può intuire che V_{OL} deve essere più piccola della più piccola tensione di soglia dei MOSFET che possono essere implementati nel progetto. Per permettere questo risultato, è necessario **forzare il MOSFET della porta logica in regione di triodo**, lì dove la caduta di tensione V_{DS} risulti essere molto bassa; tali considerazioni possono essere ritrovate osservando il grafico delle caratteristiche di uscita di un NMOS ad arricchimento, riportando anche diverse rette di carico: è evidente che, desiderando che l'intersezione tra la retta di carico e la caratteristica all'ingresso alto ($V_{IN} = V_{DD}$) cada in regione di triodo, è possibile agire sul rapporto W/L del dispositivo, in modo da **alzare la sua curva $I_d - V_{DS}$, o sulla resistenza di carico, in modo da diminuire la pendenza della retta di carico**. In realtà, questo secondo accorgimento non è produttivamente utile, dal momento in cui elevate resistenze implicano elevati tempi di propagazione.



Per quantificare le considerazioni qualitative appena fatte, è possibile scrivere una LKT sulla maglia di uscita:

$$V_{OUT} = V_{DD} - I_D R_{LOAD}$$

Mentre, per il MOSFET in regione di triodo e per $V_{GS} = V_{IN} = V_{DD}$:

$$I_D = \frac{1}{2} \mu_n C_{ox} \cdot \frac{W}{L} [2(V_{DD} - V_{th})V_{OL} - V_{OL}^2]$$

Considerando che già V_{OL} è abbastanza piccolo, il suo quadrato è praticamente trascurabile; quindi:

$$V_{OL} = V_{DD} - \mu_n C_{ox} \cdot \frac{W}{L} (V_{DD} - V_{th}) V_{OL} R_{LOAD}$$

Da cui:

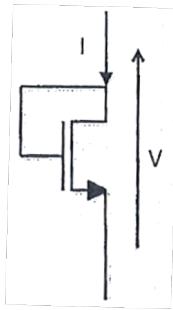
$$V_{OL} = \frac{V_{DD}}{1 + \mu_n C_{ox} \cdot \frac{W}{L} (V_{DD} - V_{th}) R_{LOAD}}$$

Si possono anche osservare gli effetti del rapporto W/L e della resistenza R_{LOAD} modellati qualitativamente in precedenza.

IL MOSFET COME BIPOLO DI CARICO

Il circuito appena mostrato è utile solo dal punto di vista didattico, nelle applicazioni reali trova poco impiego ed il motivo risiede nell'eccessivo spazio di cui necessita per funzionare in condizioni accettabili; infatti, se si volesse agire aumentando la resistenza di carico si dovrebbe agire o sulla resistività ρ , che sul silicio in elettronica integrata non riesce ad essere troppo elevata, o sul rapporto L/S , che induce una richiesta di spazio non giustificabile per i pochi vantaggi che porta con sé. Per questi motivi, quando si parla di applicazioni reali, si preferisce evitare l'impiego di dispositivi che non siano transistor (in questo caso MOSFET), in modo da ottimizzare al meglio lo spazio sul dice. Sulla base di quanto appena detto, è necessario specificare che non è preclusa la possibilità di realizzare un resistore sul silicio in elettronica integrata ma solo che la sua implementazione introduce più svantaggi (soprattutto dal punto di vista di gestione dello spazio) che vantaggi.

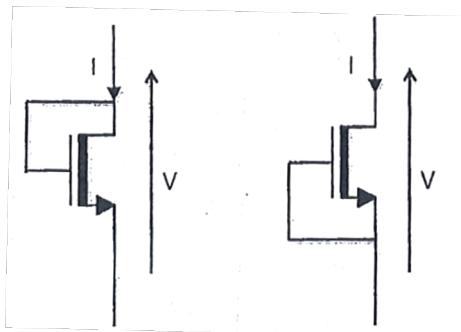
Ci si chiede, allora, come possa un transistor sostituire un resistore, soprattutto quando quest'ultimo è un bipolo a due terminali mentre il MOSFET ne ha tre (utilizzabili). Per poter trasformare un bipolo a tre terminali in un bipolo a due terminali è necessario collegarne due tra di loro, in modo da non poter accedere ad uno dei due dall'esterno; per quanto riguarda il MOSFET, esistono diverse configurazioni che permettono al dispositivo di lavorare con soli due terminali ma solo poche di queste risultano in un comportamento quasi lineare, simil resistore. Per un NMOS ad arricchimento l'unico collegamento possibile che permette il funzionamento a due terminali è il cortocircuito tra Gate e Drain:



Dal momento in cui $V_{DS} = V_{GS}$, il MOSFET si troverà sicuramente in pinch – off e la corrente di Drain sarà legata alla tensione V applicata tramite l'equazione (di pinch – off):

$$I = \frac{1}{2} \mu_n C_{ox} \cdot \frac{W}{L} (V - V_{th})^2$$

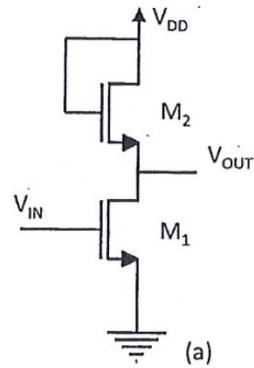
Ovviamente, se $V < V_{th}$, la corrente I sarà nulla perché il MOSFET non sarebbe conduttivo; segue che questo dispositivo può essere utilizzato solo per tensioni superiori a quella di soglia del transistore. Queste considerazioni relegano il NMOS ad arricchimento a strumento puramente didattico, come sarà più chiaro a breve.



Per quanto riguarda gli NMOS a svuotamento, si ricordi che il canale tra Drain e Source viene preformato mediante un opportuno passo tecnologico; pertanto, il funzionamento non differisce dall'omologo dispositivo ad arricchimento se non per una tensione di soglia negativa. Per tale motivo, il collegamento possibile per rendere il dispositivo a due terminali non è solo quello tra Gate e Drain, ma anche quello tra Gate e Source; il primo dei due, tuttavia, introduce le stesse problematiche del NMOS ad arricchimento e porterà a soffermarsi di più sul secondo.

Poiché il canale è già formato, la corrente potrà scorrere anche a $V_{GS} = 0$; in tal caso, il legame tra corrente I e tensione V sarà proprio la curva $I_D - V_{DS}$ del MOSFET individuata dal parametro $V_{GS} = 0$.

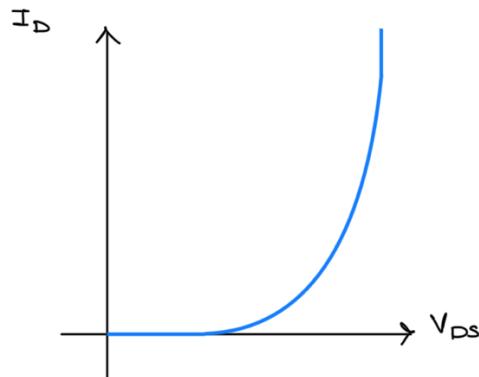
Le configurazioni di invertitore che sfruttano questi “nuovi” dispositivi sono quelle a NMOS – EE e a NMOS – ED (la prima E indica che il MOSFET interruttore è ad arricchimento). Per quanto riguarda la prima, dove la E sta per Enhancement (arricchimento), il MOSFET M_1 funge da interruttore e M_2 da dispositivo di carico:



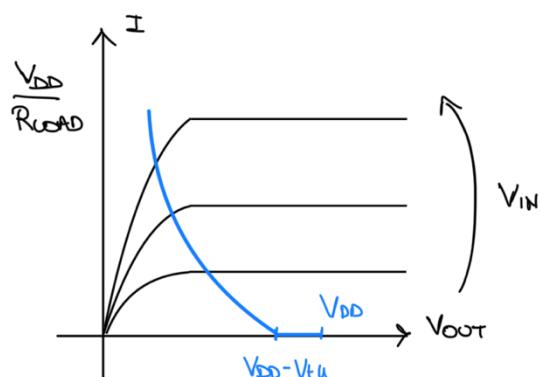
La caduta ai capi di M_2 deve essere almeno pari alla sua tensione di soglia V_{th} e, per questo motivo, la massima tensione di uscita non potrà mai essere superiore a:

$$V_{OH} = V_{DD} - V_{th}$$

La caratteristica del MOSFET di carico M_2 , in cui sono cortocircuitati Gate e Drain (e quindi $V_{GS} = V_{DS}$, individuando una sola curva), è graficato come segue:



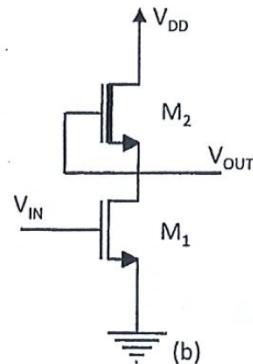
Mentre nel piano ausiliario (in cui va riflesso e traslato di V_{DD}):



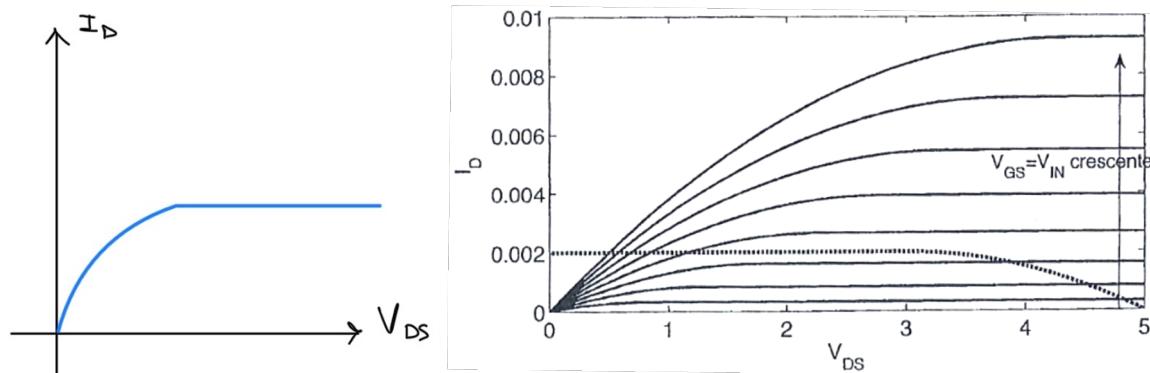
Ma in corrispondenza di ingressi bassi (e quindi correnti nulle, selezionando la curva interdetta di M_1) **la soluzione non è univoca**, c'è **un intervallo di soluzioni**; tuttavia, **una volta giunto all'estremo inferiore** di tale intervallo, **il MOSFET non carica la capacità parassita** (dovuta agli altri stadi e della quale si discuterà a breve), **stabilizzando il valore di tensione a $V_{DD} - V_{th}$** .

Ed è qui che risiede la **poca praticità di questo dispositivo**; infatti, la V_{OH} corrispondente è più **piccola di quanto modellato con gli schemi precedenti** e ciò risulta in un **minor margine di rumore**. Da queste considerazioni, si può ben intuire come sia **molto più utile e produttivo**

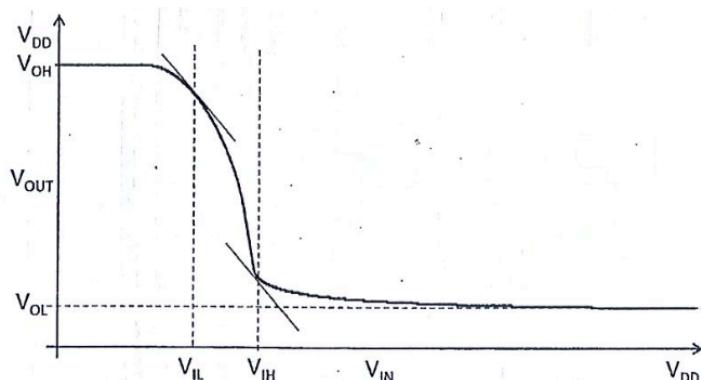
l'impiego di NMOS a svuotamento, anche perché **non introducono alcuna complicazione**, né tecnologica né progettuale.



Poiché risulta essere particolarmente utile, **questo tipo di configurazione verrà approfondita più nel dettaglio**. In figura sono mostrati i grafici relativi al MOSFET di carico M_2 e alla sovrapposizione della caratteristica $I_D - V_{DS}$ del MOSFET M_1 con la curva di carico (anch'essa caratteristica $I_D - V_{DS}$) del MOSFET M_2 ribaltata rispetto all'asse delle correnti e traslata di V_{DD} :



I punti di intersezione tra le due caratteristiche rappresentano, al variare della tensione di ingresso $V_{IN} = V_{GS}$, tutte le tensioni di uscita dell'invertitore; pertanto, è possibile costruire per punti il grafico della caratteristica di trasferimento $V_{IN} - V_{OUT}$:



Si noti una forte somiglianza con l'invertitore a resistenza di carico: il grafico può essere diviso in entrambi i casi in tre regioni distinte, in quelle alle estremità l'uscita non dipende (o dipende pochissimo) dalle variazioni di ingresso mentre quella centrale ha una pendenza tendenzialmente minore di -1.

Per valutare V_{OH} , si tenga in considerazione il fatto che **per tensioni di ingresso inferiori alla soglia di M_1 nel circuito non può scorrere corrente e la tensione di uscita dovrà necessariamente e integralmente essere la tensione di alimentazione** (quindi $V_{OH} = V_{DD}$), come nel caso dell'invertitore a carico resistivo. La valutazione di V_{OL} è leggermente più articolata, si noti che **per tensioni di ingresso alte ($V_{IN} = V_{DD}$) la curva di carico interseca le caratteristiche di M_1 in regione di triodo mentre essa è in regione di pinch – off; pertanto, basta uguagliare le rispettive espressioni della corrente di Drain:**

$$I_{D1} = \frac{1}{2} \mu_n C_{ox} \cdot \frac{W_1}{L_1} [2(V_{DD} - V_{th})V_{OL}] \wedge I_{D2} = \frac{1}{2} \mu_n C_{ox} \cdot \frac{W_2}{L_2} |V_{TD}|^2$$

Ottenendo:

$$\frac{W_1}{L_1} [2(V_{DD} - V_{th})V_{OL}] = \frac{W_2}{L_2} |V_{TD}|^2$$

Definendo:

$$k_R = \frac{W_1}{W_2} \cdot \frac{L_2}{L_1}$$

Si ottiene:

$$V_{OL} = \frac{|V_{THD}|^2}{2k_R(V_{DD} - V_{th})}$$

Per ottenere una V_{OL} quanto più bassa possibile sarà necessario dimensionare adeguatamente i MOSFET in modo tale che il rapporto delle loro transconduttanze sia più alto possibile (nonostante k_R non superi normalmente 5 – 6). Un altro modo con cui definire tale parametro è:

$$k_R = \frac{k_1}{k_2} = \frac{\frac{1}{2} \mu_n C_{ox}}{\frac{1}{2} \mu_n C_{ox}} \left(\frac{W_1 L_2}{W_2 L_1} \right) = \frac{W_1}{W_2} \cdot \frac{L_2}{L_1}$$

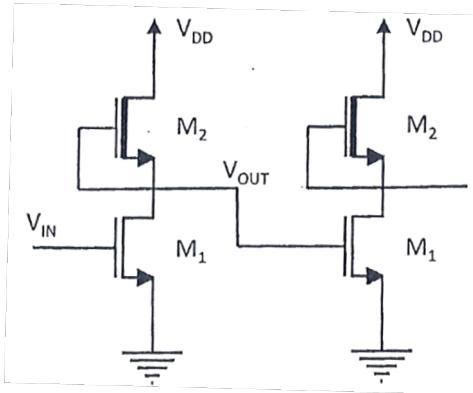
Nonostante la simbologia possa ingannare, k_R non è una transconduttanza ma un numero puro e dipende unicamente da parametri geometrici:

$$[k_R] = \frac{[k_1]}{[k_2]} = \left[\frac{V^2/A}{V^2/A} \right] = [1]$$

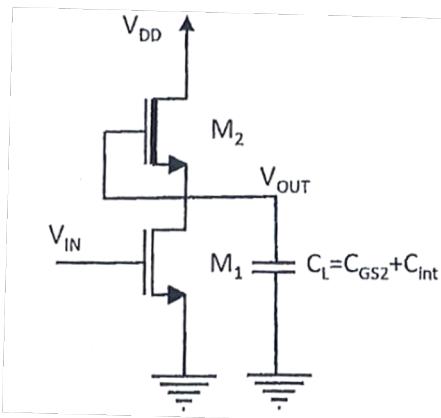
Poiché gli invertitori così realizzati determinano il loro valore di tensione nominale basso tramite il rapporto dei parametri geometrici dei MOSFET coinvolti, questo tipo di circuiti prende il nome di **logiche a rapporto**.

Un'ultima considerazione può essere fatta sul tratto apparentemente verticale della caratteristica di trasferimento dell'invertitore in esame; si noti che, in quel punto, entrambi i MOSFET sono in pinch – off e, se non si considerasse alcun effetto lineare, si potrebbe effettivamente dire che tale tratto è a pendenza perfettamente verticale. Nella pratica, però, il tratto di pinch – off ha una pendenza (seppur minima) e ciò risulta in una pendenza della caratteristica non verticale ma piuttosto ripida.

Analizzando sempre un invertitore NMOS con carico a svuotamento, si consideri il seguente circuito al fine della **valutazione dei tempi di propagazione**:

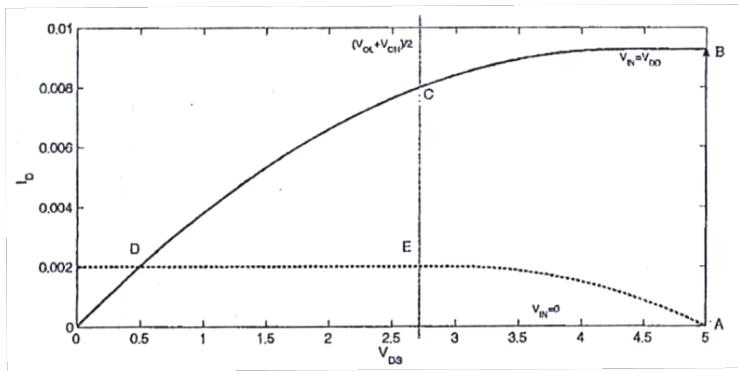


In realtà, **nella pratica ciò che è visto dinamicamente dal primo invertitore non è il secondo stadio in tutta la sua interezza, bensì un effetto capacitivo ($C_{ox}WL$) dovuto alla presenza dei due MOSFET**; pertanto, volendo generalizzare anche alla presenza di più stadi, il circuito da analizzare per una efficace valutazione del tempo di propagazione è:



Il contributo capacitivo di C_{GS2} è quello relativo ai MOSFET tra Gate e Source, mentre C_{int} è una capacità che modella in maniera semplificata il comportamento dinamico dei dispositivi (gli effetti capacitivi dipendono fortemente dalla tensione in maniera non lineare e il considerarli rigorosamente appesantirebbe solo la trattazione); queste semplificazioni rendono i MOSFET dei dispositivi istantanei: al variare della tensione segue immediatamente una variazione della corrente a causa della perdita della propria reattività. Si ricordi che quando si parla di effetti capacitivi e capacità non si sta considerando alcun bipolo capacitore, bensì gli effetti che risultano dalla presenza di situazioni in cui al variare della tensione c'è un accumulo locale di carica.

Sotto queste ipotesi, è possibile studiare intuitivamente il **tempo di propagazione** mediante il **metodo grafico**:



Quando l'uscita si trova al valore logico basso, $V_{IN} = V_{OL}$, il MOSFET M_1 si trova interdetto e l'uscita al valore logico alto; ci si trova nel punto indicato con **A**. Si supponga che l'ingresso commuti istantaneamente da V_{OL} a V_{DD} , spostandosi dal punto **A** al punto **B** (visto che sulla capacità di carico non possono esserci variazioni istantanee di tensione), a partire dal quale la capacità comincerà a scaricarsi secondo la legge:

$$i = C_L \frac{dV_{out}}{dt}$$

Ipotizzando che tale corrente di scarica sia costante e pari alla corrente di pinch – off del MOSFET M_1 (dal grafico si può osservare come la realtà non corrisponda alla semplificazione della situazione) si può integrare l'equazione separando le variabili omogenee (effettuando la convenzione dell'utilizzatore su C_L):

$$\begin{aligned} -k_1(V_{DD} - V_{th})^2 &= C_L \frac{dV_{out}}{dt} \Rightarrow dt = -\frac{C_L}{k_1(V_{DD} - V_{th})^2} dV_{out} \\ \int_0^{t_{PHL}} dt &= \int_{V_{DD}}^{\frac{V_{DD}+V_{OL}}{2}} -\frac{C_L}{k_1(V_{DD} - V_{th})^2} dV_{out} \\ t_{PHL} &= \frac{C_L}{k_1(V_{DD} - V_{th})^2} \int_{\frac{V_{DD}+V_{OL}}{2}}^{V_{DD}} dV_{out} = \frac{C_L(V_{DD} - V_{OL})}{2k_1(V_{DD} - V_{th})^2} \end{aligned}$$

Ci si trova, quindi, nel punto **C**. Il circuito evolve successivamente verso il punto **D**, la situazione stabile in cui l'uscita si trova al valore V_{OL} mentre l'ingresso è V_{DD} . Per valutare il t_{PLH} si possono fare considerazioni analoghe, solo che non si percorrerà più la linea piena da **B** a **C** ma quella tratteggiata da **D** ad **E**; ovvero, nel passaggio della tensione di ingresso da V_{DD} a V_{OL} il MOSFET M_1 si spegne e la capacità C_L viene caricata attraverso il MOSFET M_2 . Dal grafico, si può osservare come la corrente di carica della capacità sia proprio la corrente di pinch – off di M_2 , per cui è possibile direttamente dire (senza ripetere i conti):

$$t_{PLH} = \frac{C_L(V_{DD} - V_{OL})}{2k_2|V_{TD}|^2}$$

Quest'ultimo termine risulta essere tendenzialmente maggiore del primo, dal momento in cui (come è possibile apprezzare anche dal grafico) la corrente di pinch – off (da cui dipendono entrambi i tempi di propagazione) è decisamente maggiore nel MOSFET interruttore M_1 che nel MOSFET di carico M_2 ; ciò è dovuto al fatto che quest'ultimo, per garantire valori di V_{OL} sufficientemente bassi, debba avere un valore di k_R maggiore di 1.

In entrambi i casi, si noti che **il tempo di propagazione dipende linearmente dall'effetto capacitivo indotto dagli stadi in cascata all'uscita**; di conseguenza, **non conviene pilotare con la stessa uscita troppe porte logiche**, altrimenti aumenterebbe il parametro C_L e il circuito risulterebbe troppo lento.

Per quanto riguarda la **potenza dissipata**, è possibile riprendere la plausibile **approssimazione alla potenza dissipata staticamente** e considerare che **quando l'ingresso è al valore logico basso non può circolare corrente**, mentre **quando l'ingresso è alto ci si trova nel punto D del grafico precedente**, con **M_1 in triodo e M_2 in pinch – off**; pertanto la **corrente al circuito sarà imposta dal MOSFET di carico M_2** , visto che **in pinch – off si comporta da generatore controllato ideale**, per cui è possibile scrivere:

$$P_D \approx P_{DC} = \frac{P_{DCH} + P_{DCL}}{2} = \frac{V_{DD}k_2|V_{TD}|^2}{2}$$

La media viene fatta perché **si considerano equiprobabili le due uscite**; altrimenti, **a rigore, il calcolo della probabilità sarebbe**:

$$P_D \approx P_{DC} = P[V_{OUT} = V_{OH}]P_{DCH} + P[V_{OUT} = V_{OL}]P_{DCL}$$

NAND E NOR IN LOGICA A RAPPORTO E L'INVERTITORE CMOS

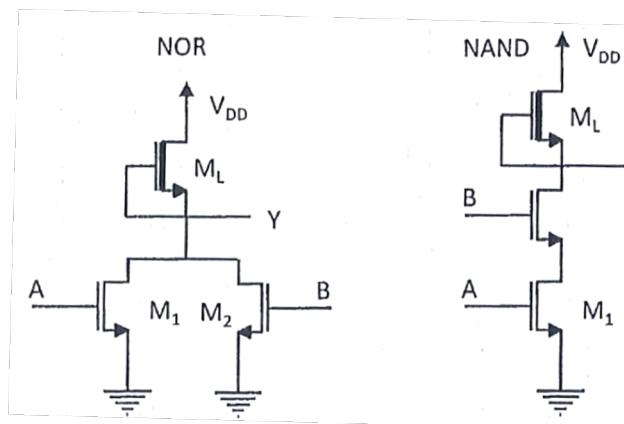
Arrivati a questo punto, **ci si chiede l'importanza** (oltre a quella didattica) **di un invertitore per la creazione di circuiti digitali che convertano funzioni logiche in porte logiche funzionali**. Il motivo risiede **nell'importanza delle porte logiche NAND e NOR**; infatti, **a differenza delle corrispettive porte non negate, AND e OR** (più semplici da realizzare), **queste due rappresentano un insieme funzionalmente completo**, ovvero un **set di funzioni che in accoppiamento con tutte le leggi dell'algebra booleana permettono di realizzare qualsiasi funzione logica di possa pensare**. Questa proprietà **permette a sole due porte di codificare un insieme infinito di operazioni**, garantendo una **enorme versatilità** e una **buona ottimizzazione dello spazio sul silicio**, al costo di **un leggerissimo** (e sostenibile) **incremento della complessità di progettazione**.

Per realizzare dal punto di vista circuitale le porte logiche in questione si parte dal presupposto che esse rappresentano la negazione della AND e della OR:

X	Y	X AND Y	X	Y	X NAND Y
0	0	0	0	0	1
0	1	0	0	1	1
1	0	0	1	0	1
1	1	1	1	1	0

X	Y	X OR Y	X	Y	X NOR Y
0	0	0	0	0	1
0	1	1	0	1	0
1	0	1	1	0	0
1	1	1	1	1	0

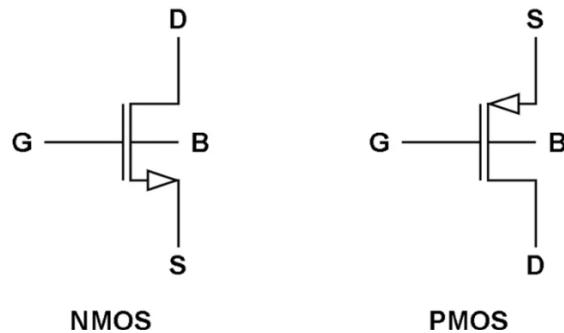
Osservando prima la **porta OR**, si può intuire come essa sia **alta solo se X, Y o entrambi sono alti**, e, quindi, potrà essere schematizzata come un **parallelo tra i segnali logici X e Y**; infatti, **tra i nodi del parallelo sussiste una tensione non nulla solo se almeno su uno dei due rami sussiste una tensione non nulla**. Analogamente per la **porta AND**, si può intuire come essa sia **alta solo se X e Y sono alti** e, quindi, potrà essere schematizzata come una **serie tra i segnali logici X e Y**; infatti, **tra i nodi della serie sussiste una tensione non nulla solo se su entrambi i rami della serie sussiste una tensione non nulla**. A partire da queste porte logiche elementari, serie e parallelo di due segnali, si inserisca l'uscita in ingresso ad un invertitore per ottenere le porte **NAND** e **NOR**:



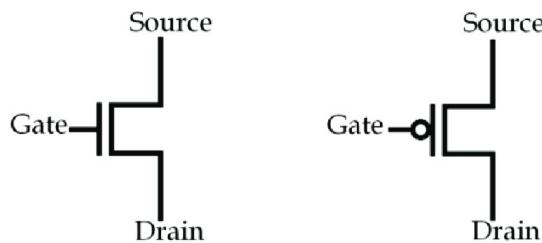
Dal punto di vista progettuale, i due collegamenti dei segnali non sono equivalenti; infatti, in parallelo i MOSFET M_1 e M_2 saranno pilotati dalle stesse tensioni ma avranno due correnti di Drain diverse (viceversa per la serie). Di conseguenza, la corrente totale che scorre in una OR a N ingressi è N volte la corrente del singolo dispositivo (che corrisponde ad un incremento di un fattore di N della larghezza del canale, NW), mentre la caduta di tensione in una AND a N ingressi è N volte la caduta di tensione ai capi del singolo dispositivo (che corrisponde ad un incremento di un fattore di N della lunghezza del canale, NL).

In un periodo cronologicamente successivo all'introduzione di queste schematizzazioni, furono sviluppate anche delle logiche di realizzazione dei circuiti digitali basati sull'impiego di MOSFET a canale P (ingegnerizzati solo molto tempo dopo rispetto agli NMOS). La tecnologia che si vuole approfondire è detta **CMOS**, **Complementary MOS**, e, sebbene il nome possa ricondurre agli NMOS o PMOS, non si parlerà di dispositivi ma di logiche di realizzazione; infatti, seguendo l'andazzo delineato dal MOSFET, la lettera antecedente al tipo di dispositivo dovrebbe indicare quale carica libera è coinvolta nella conduzione ma, a questo istante, non si conosce alcun tipo di carica libera indicata con C e, di conseguenza, si può dedurre che CMOS non delinea un nuovo dispositivo.

Prima di addentrarsi nello studio degli invertitori (e delle porte logiche) in logica CMOS, si vuole fare una piccola digressione sulla simbologia e sul funzionamento di questi dispositivi. I PMOS e gli NMOS sono, a rigore, indicati con la seguente simbologia:



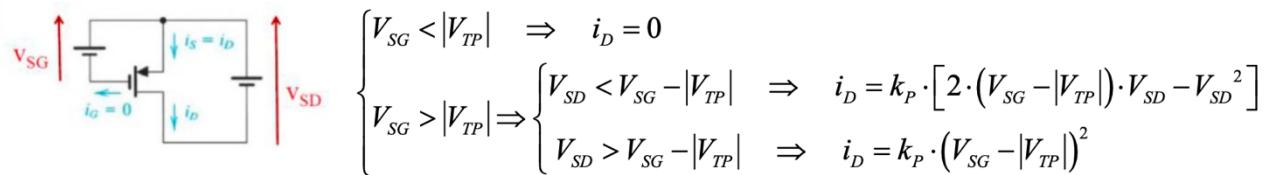
Tuttavia, come è stato già anticipato discutendo dei MOSFET, **si parla di dispositivi perfettamente simmetrici** e (a differenza del BJT) **la determinazione della posizione della freccia non è da associare ad una caratteristica intrinseca del dispositivo** (dov'era il drogaggio maggiore/minore) ma alla sua collocazione nel circuito; infatti, **la freccia va sempre posizionata sul Source, che è il terminale a potenziale maggiore** (NNOS) o minore (PMOS). Per evitare di dover determinare sempre quale è il potenziale maggiore o minore, e solo successivamente trovare la posizione e il verso della freccia, **si può introdurre una nuova simbologia** (da contestualizzare prevalentemente all'elettronica digitale) che fa uso del cosiddetto “**pallino di negazione**”, **tipico delle porte invertenti, per specificare il PMOS** e che elimina qualsiasi distrazione prodotta dalle frecce:



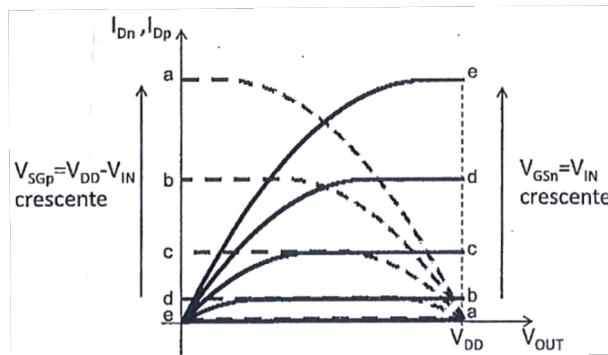
Sebbene le figure a seguire non disporranno di questa simbologia (per una questione di riferimenti al testo) si tenga conto che la conversione dall'una all'altra è sempre possibile considerando:

- Il “pallino di negazione” laddove la freccia punta verso il gate;
- L’assenza di frecce laddove la freccia punta lontano dal gate.

Si vuole, adesso, riportare la **caratteristica di trasmissione di un PMOS**, omessa in precedenza:



Graficamente (in relazione al NMOS, indicato con la linea piena):

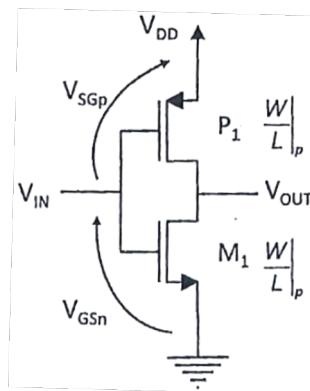


Supponendo:

$$k_N = k_P \wedge |V_{th}| = |V_{TP}|$$

I due grafici sono perfettamente uguali se non per un fattore di riflessione degli assi (infatti, per il PMOS si parla di $V_{SG} = -V_{GS}$ e simili).

Di seguito è proposto lo **schema realizzativo di un invertitore CMOS**, composto da **due MOSFET**: il MOSFET “di carico” è a canale P, mentre quello “di segnale” è a canale N; i due **condividono i terminali di Drain** (su cui è posta l’uscita V_{OUT}) e di Gate (su cui è posto l’ingresso V_{IN}), mentre l’alimentazione è posta sul Source del PMOS e la massa al Source del NMOS. Il tutto si rappresenta come segue:



Sebbene siano stati indicati così, **per questo tipo di interruttore non è possibile fare la distinzione tra dispositivo di carico e dispositivo di segnale**; infatti, come si è potuto osservare, **sia l’ingresso che l’uscita sono sui terminali comuni ad entrambi i MOSFET**:

$$V_{GSn} = V_{IN} \wedge V_{SGp} = V_{DD} - V_{IN}$$

Si può, quindi, intuire che **l’aumento progressivo della tensione di ingresso V_{IN} porta in conduzione il NMOS e spegne progressivamente il PMOS, e viceversa; i due dispositivi funzionano in maniera complementare** (C sta per Complementary).

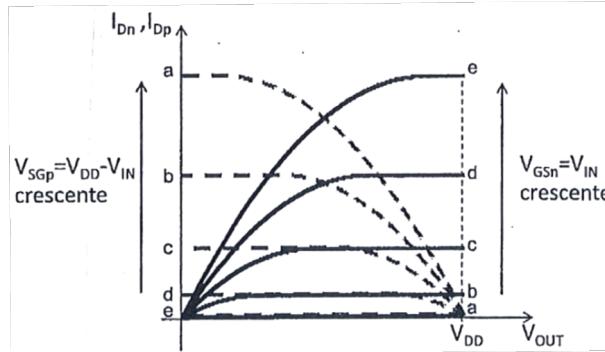
L’analisi del circuito alle due condizioni di funzionamento estreme (statiche) $V_{IN} = V_{DD}$ e $V_{IN} = 0$ porta a fare considerazioni notevoli; in primis si noti che, **nel primo caso, il NMOS è in conduzione e il PMOS è interdetto**, mentre **nel secondo caso accade il contrario**, inducendo a pensare che **nei punti di funzionamento statico ci sarà sempre un dispositivo non conduttivo e, quindi, una corrente di Drain inevitabilmente nulla**. Le conseguenze di queste considerazioni sono due:

1. L'escursione logica dell'invertitore è pari alla tensione di alimentazione;
 - a. $V_{OUT} = V_{DD}$ nel primo caso e $V_{OUT} = 0$ nel secondo;
2. La potenza dissipata staticamente (e quindi ragionevolmente quella totale) è nulla;
 - a. Non c'è scorIMENTO di corrente.

Per studiare la **caratteristica di trasferimento** di questo tipo di invertitore **si analizzi simultaneamente il comportamento dei due dispositivi complementari**, supponendo che:

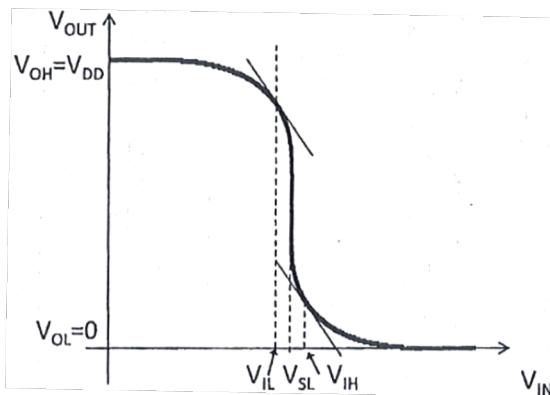
$$\mu_n \frac{W_N}{L_N} = \mu_p \frac{W_P}{L_P} \wedge |V_{th}| = |V_{TP}|$$

Il grafico delle due caratteristiche $I_D - V_{DS}$ è uguale a quello precedentemente mostrato:



Non essendoci più una sola curva di carico, è necessario etichettare ogni singola curva sulla base delle singole condizioni di funzionamento, in modo tale da poter accoppiare più facilmente le singole curve caratteristiche dei dispositivi e determinare il punto di funzionamento V_{OUT} sulla base della tensione V_{IN} di ingresso.

Senza considerare l'effetto di modulazione della lunghezza di canale per una questione di semplicità, si noti che in corrispondenza delle lettere c, d ed e non si ha un unico punto ma un intervallo di intersezioni, visto che entrambi i MOSFET si ritrovano in pinch-off; questi tratti sono indicativi di una pendenza piuttosto elevata (e, quindi, di una demarcata V_{SL}) nella caratteristica dell'invertitore, che può essere individuata dal grafico seguente:



In corrispondenza di questa regione, come è stato anticipato, **entrambi i dispositivi si trovano in pinch-off** e, per determinare V_{SL} , **non si fa altro che ugualare le relative correnti di Drain**:

$$I_{Dn} = I_{Dp}$$

$$k_N(V_{IN} - V_{th})^2 = k_P(V_{DD} - V_{IN} - |V_{TP}|)^2$$

$$k_N(V_{SL} - V_{th})^2 = k_P(V_{DD} - V_{SL} - |V_{TP}|)^2$$

$$\sqrt{k_N}(V_{SL} - V_{th}) = \sqrt{k_P}(V_{DD} - V_{SL} - |V_{TP}|)$$

$$V_{SL} = \frac{\sqrt{\frac{k_P}{k_N}}(V_{DD} - |V_{TP}|) + V_{th}}{1 + \sqrt{\frac{k_P}{k_N}}}$$

Ricordando che V_{SL} è definita nel punto in cui $V_{IN} = V_{OUT} = V_{SL}$. Volendo esprimere il tutto in termini di $k_R = k_N/k_P$:

$$V_{SL} = \frac{V_{DD} - |V_{TP}| + \sqrt{k_R}V_{th}}{1 + \sqrt{k_R}}$$

Si osservi che, nel caso in cui $k_N = k_P$ e $|V_{th}| = |V_{TP}|$:

$$V_{SL} = \frac{V_{DD}}{2}$$

L'invertitore, sotto queste particolari ipotesi, viene detto simmetrico, mentre i MOSFET che lo compongono sono detti simmetrizzati. L'invertitore simmetrizzato è **particolarmente utile** (perché la sua caratteristica da $V_{DD}/2$ a V_{DD} è pari a quella tra 0 e $V_{DD}/2$ ribaltata e perché i margini di rumore sono uguali) e **può essere ottenuto dimensionando in questo modo i MOSFET:**

$$L_N = L_P \wedge W_P = \frac{\mu_n}{\mu_p} W_N$$

Infatti, **vanno uguagliati i fattori di transconduttanza** (da cui segue anche l'uguaglianza e $|V_{th}| = |V_{TP}|$):

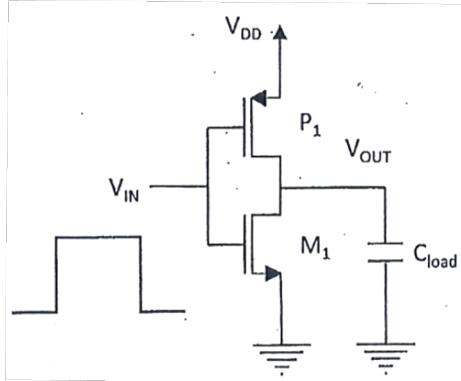
$$\begin{aligned} \mu_n C_{ox} \cdot \frac{W_N}{L_N} &= \mu_p C_{ox} \cdot \frac{W_P}{L_P} \\ \mu_n \frac{W_N}{L_N} &= \mu_p \frac{W_P}{L_P} \Rightarrow \begin{cases} \mu_n W_N = \mu_p W_P \\ L_N = L_P \end{cases} \end{aligned}$$

C_{ox} è lo stesso perché questi invertitori sono sempre realizzati in elettronica integrale (infatti $\mu_n C_{ox}/2$ è chiamata transconduttanza di processo) e il parametro in questione **dipende dalle condizioni del silicio da cui i MOSFET sono ricavati**.

Dalle condizioni di simmetrizzazione, in particolare quelle sul dimensionamento dei MOSFET, si intuisce che, **a parità di lunghezza del canale, il PMOS è più largo, aumentando i tempi di propagazione ma diminuendo la corrente di conduzione**.

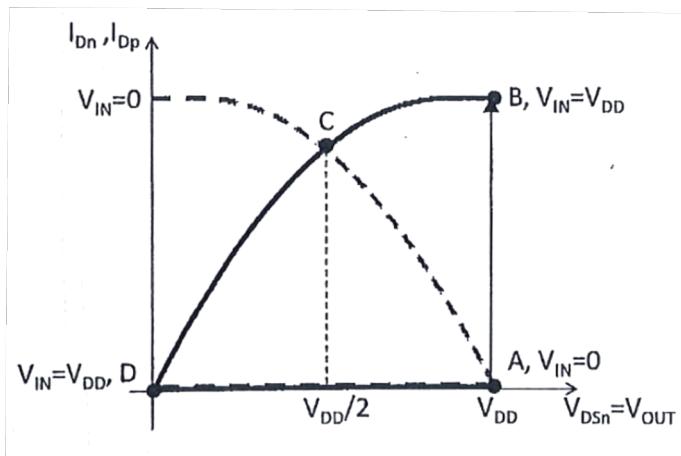
Ciò che rimane da analizzare per avere un quadro chiaro degli invertitori in logica CMOS sono le **proprietà dinamiche: tempo di propagazione e potenza dissipata** (è una proprietà dinamica perché, come è stato dimostrato, la potenza staticamente dissipata è nulla).

Il circuito utilizzato per studiare il tempo di propagazione è il seguente, sapendo che C_L è la capacità che ingloba gli effetti capacitivi di tutti gli stadi potenzialmente collegati in serie all'invertitore e le capacità interne degli NMOS e PMOS, delle quali è trascurata la dipendenza dalla tensione applicata; sotto queste ipotesi, **l'invertitore CMOS è un circuito in cui variazioni di tensione istantanee portano variazioni di corrente istantanee**.



Applicando in ingresso un impulso rettangolare di tensione di durata opportuna e con fronti di salita istantanei, si valuti in prima istanza il passaggio da valore logico basso a valore logico alto:

$$V_{IN} = 0V \rightarrow V_{IN} = V_{DD}$$



Sulla capacità di carico, inizialmente carica alla tensione V_{DD} , si porterà a $V_{DD}/2$ in un tempo t_{PHL} determinato dalla relazione:

$$i = C_L \frac{dV_{OUT}}{dt}$$

Nell'istante immediatamente precedente la commutazione ci si trova nella condizione indicata con il punto A. Successivamente, il MOSFET N entra in conduzione e P in interdizione; non potendo variare l'uscita a causa della capacità, ci si deve spostare direttamente nel punto B, a partire dal quale la capacità inizierà a scaricarsi muovendosi verso il punto C secondo la curva caratteristica di N. In corrispondenza del punto C la valutazione del t_{PHL} termina, dal momento in cui la tensione raggiunta è proprio $V_{DD}/2$; tuttavia, il circuito completerà la propria commutazione portandosi nel punto di funzionamento stabile indicato con D.

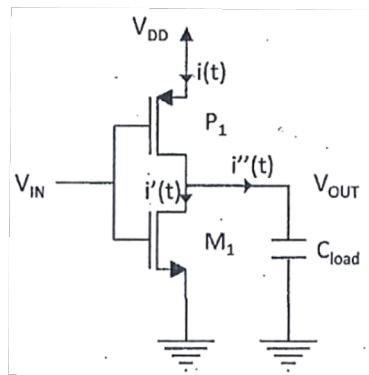
Utilizzando l'espressione della **corrente di pinch – off del MOSFET N** , è possibile valutare il t_{PHL} come segue:

$$\int_0^{t_{PHL}} dt = \frac{C_L}{k_N(V_{DD} - V_{th})^2} \int_{\frac{V_{DD}}{2}}^{V_{DD}} dV_{OUT} \Rightarrow t_{PHL} = \frac{C_L V_{DD}}{2k_N(V_{DD} - V_{th})^2}$$

Poiché l'invertitore CMOS è simmetrico, si può concludere l'analisi del tempo di propagazione considerando che:

$$t_{PLH} = t_{PHL}$$

Utilizzando lo stesso schema della figura precedente, **si separi la corrente assorbita dall'alimentazione $i(t)$ in due componenti, $i'(t)$ e $i''(t)$** , scrivendo un'equazione di Kirchhoff alle correnti sul nodo di uscita:



Si riconosce in $i''(t)$ la corrente che serve a caricare la capacità C_L (la scarica avviene senza dissipazione di potenza dell'alimentazione), mentre si ipotizza che $i'(t)$ sia ciò che serve ai dispositivi per cambiare internamente il loro stato. Di conseguenza, la potenza dinamicamente dissipata può essere separata in due contributi:

$$P_D = P'_D + P''_D$$

In questa sede **si vorrà valutare solo qualitativamente P''_D per una questione di semplicità**; infatti, a causa della complessità delle equazioni che modellano i transitori, una soluzione quantitativa globale sarebbe possibile solo ricorrendo a simulatori circuituali. In primis, **si supponga che tutte le capacità interne dei dispositivi attivi siano inglobate nelle capacità di carico** (implicando automaticamente la risposta istantanea dei MOSFET alle variazioni di tensione); inoltre, **si ipotizzi che le due aliquote di potenza dissipata siano tra di loro indipendenti**, in modo da poter essere valutate separatamente.

La potenza sotto esame, P''_D , corrisponde alla potenza media dissipata per caricare, in un periodo, la capacità C_L ; pertanto, si avrà:

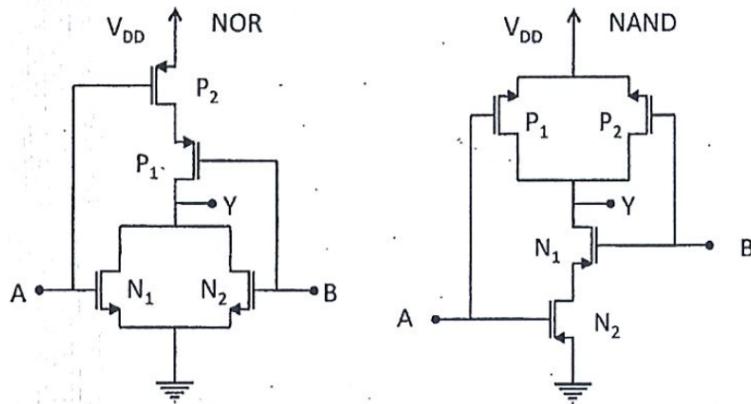
$$P''_D = \frac{1}{T} \int_T^\square v(t)i''(t)dt = \frac{C_L}{T} \int_T^\square v(t) \frac{dv(t)}{dt} dt = \frac{C_L}{T} \int_0^{V_{DD}} v dv = \frac{C_L V_{DD}^2}{T} = f C_L V_{DD}^2$$

Il risultato a cui si è pervenuti rappresenta, probabilmente, la legge più importante a cui hanno obbedito i circuiti elettronici digitali negli ultimi anni; in particolare, permette di osservare la dipendenza della potenza dissipata dalla frequenza a cui il circuito lavora. Con l'avanzamento tecnologico, sono aumentate notevolmente le frequenze di clock dei moderni microprocessori

(arrivando anche nell'ordine dei GHz) e si è cercato di compensare riducendo le tensioni di alimentazione (che hanno un impatto quadratico sulla potenza dissipata) e le capacità di carico (tramite un opportuno dimensionamento dei dispositivi sul silicio che ha evidenziato ancora di più l'importanza della miniaturizzazione).

PORTE E FUNZIONI LOGICHE COMPLESSE

Volendo dapprima analizzare le **porte logiche elementari in logica CMOS**, si noti che oltre al classico collegamento in serie e in parallelo specificato per la logica a rapporto, è **necessario implementare anche una rete di PMOS collegati in maniera complementare alla parte di segnale** (in serie per la NOR e in parallelo per la NAND):



Prendendo in esame la NOR, si osservi che l'uscita Y è al valore logico alto se A o B (o entrambi) si trovano al valore logico alto; osservando la rete PMOS, Y potrà portarsi al valore logico basso se è scollegata dall'alimentazione, ovvero se A, B o entrambi sono tali dal portare in interdizione almeno uno dei PMOS, ovvero se si trovano al valore logico alto e i dispositivi sono connessi in serie. Le stesse considerazioni possono essere fatte per la NAND e, come sarà più chiaro a breve, sono valide per qualsiasi porta logica complessa CMOS: esisterà una rete NMOS con dispositivi in serie e in parallelo (rete di pull – up) e una rete PMOS costruita in maniera duale (rete di pull – down).

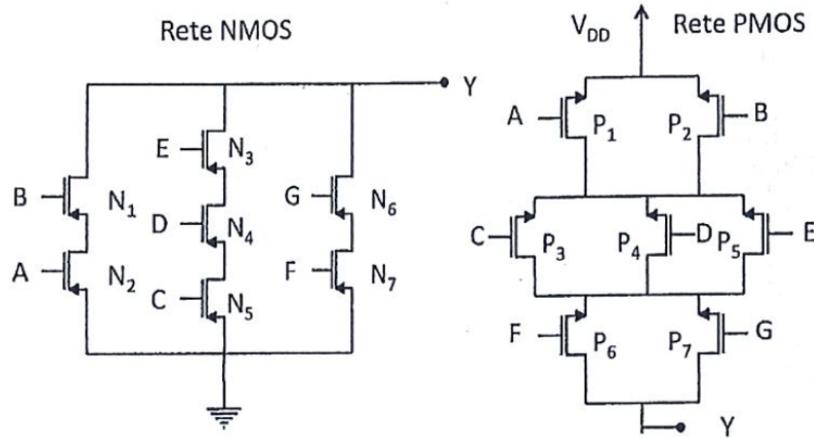
Per valutare quale logica di collegamento (serie o parallelo) è più conveniente e dove, è necessario completare la discussione sulla logica CMOS andando a sviluppare determinate considerazioni (quantomeno qualitative). A differenza della tecnologia NMOS, in cui erano presenti solo dispositivi a canale N, in questo caso sono presenti dispositivi caratterizzati da entrambe le polarità possibili; inoltre, in entrambi i circuiti sono presenti sia collegamenti in serie che in parallelo. Ci si chiede quale sia il miglior parametro elettrico da tenere in considerazione per valutare il confronto: poiché i margini di rumore sono indipendenti dalla geometria dei dispositivi, non rimane che effettuare il confronto a parità di tempi di propagazione.

Dal momento in cui i dispositivi a canale P sono già penalizzati da una ridotta mobilità, al fine di mantenere il loro fattore k_p equivalente al corrispettivo N (in modo da avere tempi di propagazione simmetrici) conviene scegliere, tra le famiglie logiche, quella in cui i dispositivi a canale P sono connessi tra loro in parallelo; perciò, la famiglia di porte logiche da preferire è quella NAND.

Si vuole concludere la trattazione con un semplice esempio, ovvero andando a tradurre in circuito digitale la seguente funzione logica a 7 ingressi:

$$Y = \overline{AB + CDE + FG}$$

La funzione può essere realizzata utilizzando **tre livelli logici in cascata**: **prima la AND (segnali in serie)** **poi la OR (segnali in parallelo)** e **poi la NOT (segnali in un invertitore)**; questo approccio, sebbene sia possibile, **non è il più pratico**, a causa di **evidenti sprechi di spazio e di dispositivi**. La **logica CMOS viene in aiuto** perché permette di realizzare il **tutto in un solo livello logico**:

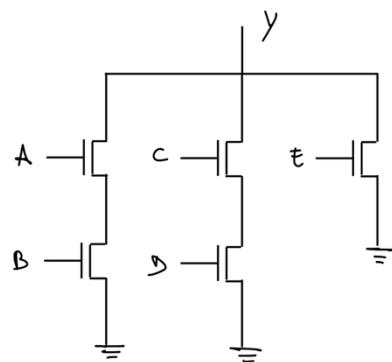


In particolare, si costruisca una rete NMOS seguendo la semplice regola di sostituzione di una AND ad una serie e di una OR ad un parallelo, per poi costruire la rete PMOS in maniera duale: laddove si vedono dispositivi in parallelo li si metta in serie e viceversa. Mettendo, poi, in comune l'uscita Y delle due reti, si realizza la funzione logica specificata.

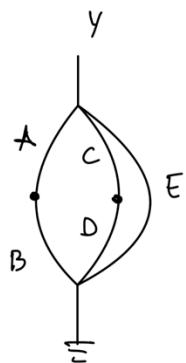
Per la realizzazione della rete di PMOS (che a breve verrà chiamata definitivamente rete di pull – up) si può utilizzare una rappresentazione grafica che prende il nome di **grafo ad archi**. Si consideri la rete di NMOS (che a breve verrà chiamata definitivamente rete di pull – down) realizzata a partire dalla funzione logica in specifica e si costruisca un grafo con una serie di archi (ciascuno relativo ad uno degli NMOS e alle cui estremità sono associati il Drain e il Source); ci sarà un solo nodo corrispondente alla massa e uno al terminale di uscita. Il grafo ottenuto sarà la rappresentazione grafica della rete di pull – down; senza ispezione visiva del circuito, per ottenere il grafo della rete di pull – up si considerano due nodi esterni e poi un nodo per ogni spazio chiuso che il grafo di pull – down individua. Tali nodi vanno, poi, uniti tra loro in modo che gli archi prodotti intersechino ogni arco del grafo di pull – down. Ad esempio, la funzione logica:

$$Y = \overline{AB + CD + E}$$

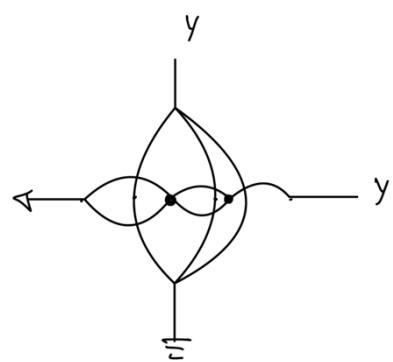
Viene resa dalla seguente rete di pull – down:



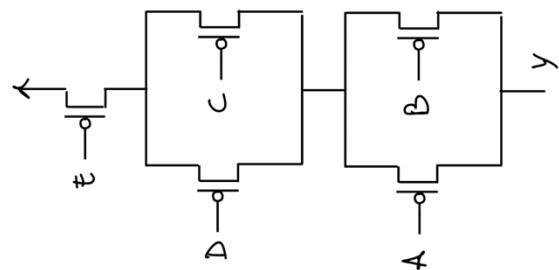
Che, tradotta in grafo ad archi, restituisce la seguente rappresentazione grafica:



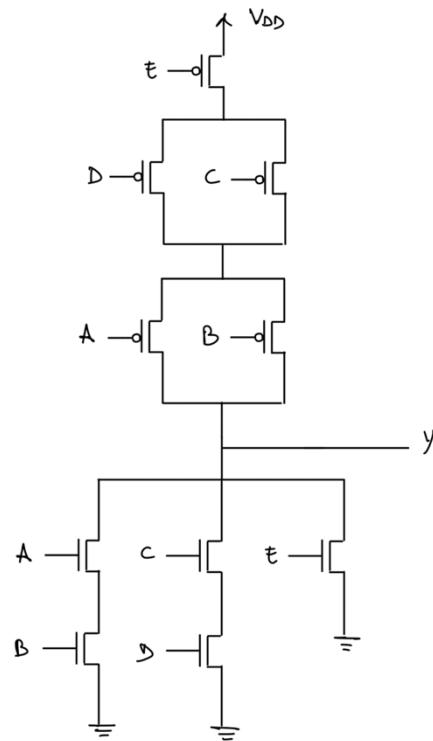
Passando poi alla rappresentazione del grafo della rete di pull – up:



Che, infine, diventa:



Ottenendo così la porta logica corrispondente in logica CMOS:

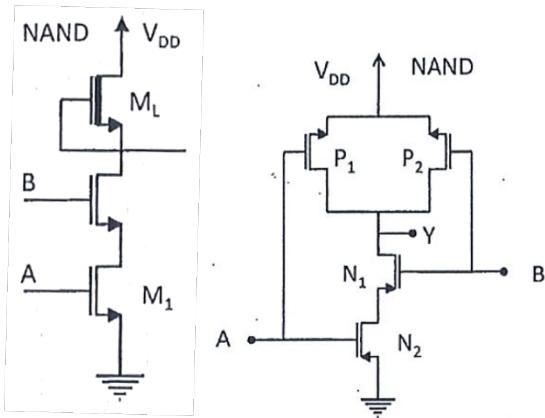


Si vuole concludere menzionando quelli che sono **i vantaggi e gli svantaggi delle due logiche di realizzazione finora considerate:**

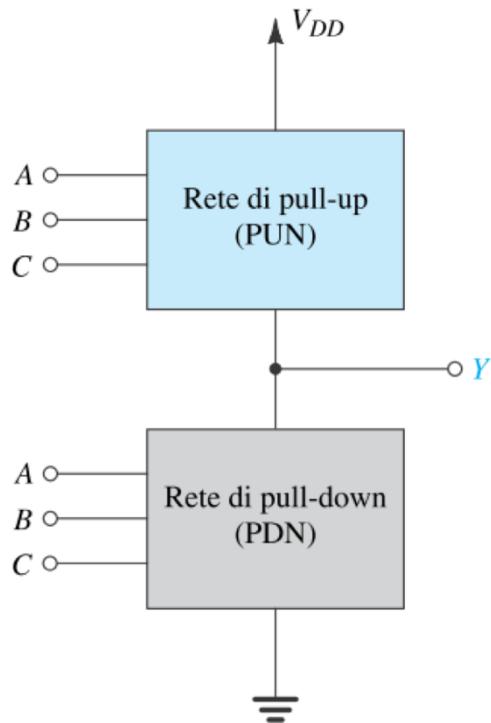
Logica a rapporto	Logica CMOS
NM non simmetrico	NM simmetrico
Tempi di propagazione diversi	Tempi di propagazione uguali
Potenza dissipata maggiore	Potenza dissipata unicamente dinamica
Minor numero di MOSFET richiesti	Maggior numero di MOSFET richiesti

Volendosi soffermare su quest'ultimo confronto, **se si dovesse realizzare una porta logica a n ingressi, in logica a rapporto saranno sempre necessari $n + 1$ dispositivi (n segnali di ingresso e un NMOS – D), mentre in logica CMOS saranno necessari n NMOS per i segnali di ingresso e n PMOS.**

Ad esempio, **per una semplice NAND a due ingressi**, in logica a rapporto sono necessari 3 MOSFET e in logica CMOS 4:

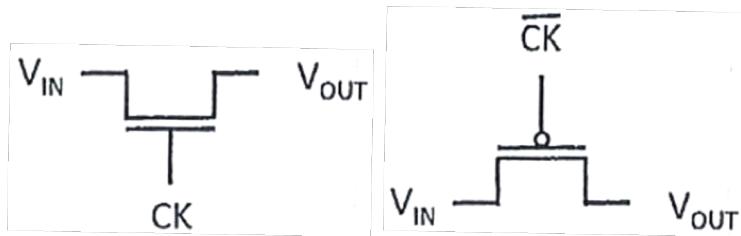


Per rete di pull – up si intende il circuito che collega l’alimentazione all’uscita e che alza il valore logico di uscita quando l’ingresso è aperto, mentre per rete di pull – down si intende il circuito che collega la massa all’uscita e che abbassa il valore logico di uscita quando l’ingresso è chiuso.



Sebbene nel corso dell’intera trattazione questa nomenclatura non sia stata utilizzata più di tanto, è importante considerare il suo impiego quando si vogliono descrivere dei circuiti per porte logiche più complesse; ad esempio, tornando al **confronto tra logiche a rapporto e CMOS**, la rete di pull – up della prima sarà costituita da un solo NMOS – D, mentre per la seconda sarà la rete complementare a quella di pull – down.

Finora i MOSFET sono stati utilizzati in elettronica digitale per cortocircuitare il nodo di uscita di una porta logica verso massa o verso l’alimentazione, quando in realtà possono essere utilizzati anche come veri e propri interruttori che consentono o meno il passaggio di un segnale. Utilizzati in questo modo, i MOSFET prendono il nome di porte di trasmissione (o pass transistors) e sono di due tipi:

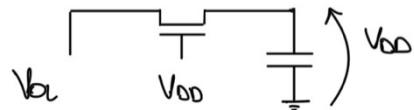


Qui ancor più che mai è **superflua la specifica della freccia per il terminale di Source**, visto che **non è noto a priori se $V_{IN} > V_{OUT}$ o viceversa e, pertanto, il terminale a potenziale maggiore non è a priori determinato**. Si noti che, in entrambi i dispositivi, se il segnale di clock (CK) è alto (ovvero a V_{DD}), il passaggio della corrente (e quindi del segnale) dal terminale di ingresso a quello di uscita è permesso e la porta di trasmissione è detta aperta.

Questi dispositivi, così come qualsiasi altra porta logica in elettronica digitale, **hanno lo scopo di essere collegati in cascata con altre reti digitali**; pertanto, **in parallelo vi verrà considerato l'agglomerato di tutti gli effetti capacitivi potenzialmente presenti**. Si consideri il pass transistor NMOS e si analizzi il passaggio di un segnale digitale alto e basso:

- **Trasmissione 0**

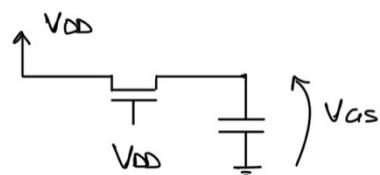
Si ha $V_{IN} = V_{OL}$ (potenzialmente 0) e la rete si traduce come segue:



Si noti che il source si sposta sul terminale di segnale (è quello a potenziale minore), portando $V_{GS} = V_{DD}$; segue che, in concomitanza con un clock alto, la capacità parassita inizia a scaricarsi, portando l'uscita a 0.

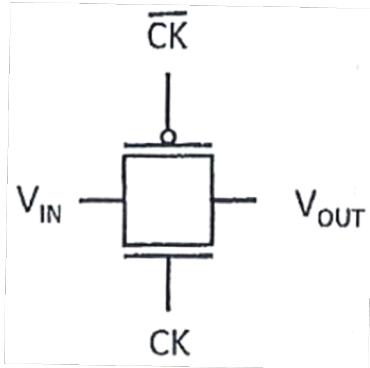
- **Trasmissione 1**

Il source, in questo caso, è vicino alla capacità, essendo l'ingresso collegato all'alimentazione:



Segue che $V_{GS} = V_{DD} - V_{OUT}$, sapendo che quest'ultimo termine tende ad aumentare; tuttavia, per $V_{GS} < V_{th}$ il MOSFET è spento e la capacità non si carica più, consentendo al dispositivo di trasmettere al più $V_{OUT} = V_{DD} - V_{th}$, che è diverso dal valore logico 1.

Come si è potuto intuire, **implementando un pass transistor in tecnologia NMOS, la trasmissione del valore logico alto subisce delle penalizzazioni** che conducono ad un **malfunzionamento del dispositivo**; dualmente, **in tecnologia PMOS, i pass transistor non garantiscono un'adeguata trasmissione del valore logico basso**. Entrambi i dispositivi, isolati, **non assolvono bene al proprio dovere** ma, per avere una porta di trasmissione quanto più funzionale possibile, **è possibile impiegare uno schema che impiega entrambe le tecnologie**:



In questo modo, quando si cerca di trasmettere un valore logico su cui una porta è debole, ci sarà sempre la complementare ad eseguire l'operazione correttamente; questo circuito prende il nome di **pass transistor in logica CMOS**.

Come si può notare, è presente sia il segnale di clock che il suo negato; sebbene in tecnologia PMOS il clock era indicato nella sua forma negata, in realtà non è mai stato necessario invertirlo, visto che era impiegato isolatamente, ma è servito solo per evidenziare il comportamento complementare al NMOS. In logica CMOS, la concomitanza di questi due segnali rende necessario l'impiego di un invertitore, aumentando il numero di MOSFET da utilizzare nella rete.

Volendo specificare il **numero di porte per ogni tecnologia**, si consideri la seguente tabella (verrà analizzato il caso peggiore, in cui il clock del PMOS va invertito con un'apposita porta logica):

Tecnologia	Numero di porte
NMOS	1
PMOS	3
CMOS	4

Ovviamente, per una questione di ottimizzazione, può essere conveniente l'utilizzo del NMOS quando il circuito “può non funzionare bene”, mentre l'utilizzo del PMOS è fortemente sconsigliato (non fa funzionare bene il circuito e richiede dispositivi in più).

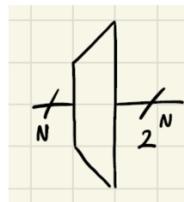
CIRCUITI COMBINATORI: DI DECODIFICA E DI INSTRADAMENTO

Per circuito combinatorio si intende un sistema, composto da un numero arbitrario di porte logiche, che prende in ingresso **N** segnali digitali e restituisce **M** segnali digitali in uscita, i quali dipendono unicamente dall'ingresso in ogni istante di tempo e sono indipendenti da alcun tipo di stato interno. Dal punto di vista unicamente formale, un circuito combinatorio è un sistema ingresso – uscita individuato da una legge del tipo:

$$Y_i = f_i(X_1, X_2, \dots, X_n) \quad \forall i \in [1, M]$$

E cioè **una legge che lega una N – upla di ingressi ad una M – upla di uscite**. I circuiti combinatori che verranno analizzati nel dettaglio in questa sede sono i **circuiti di codifica e decodifica e i circuiti di instradamento dei segnali**.

Viene definito **decoder $N - 2^N$** un particolare **circuito combinatorio** che prende N ingressi e restituisce 2^N uscite, il cui legame funzionale è di semplice comprensione: **in funzione delle 2^N possibili combinazioni degli ingressi, viene abilitata al valore logico alto una sola delle possibili uscite**.



La tabella di verità di un decoder $3 - 8$ è la seguente:

x_1	x_2	x_3	y_1	y_2	y_3	y_4	y_5	y_6	y_7	y_8
0	0	0	1	0	0	0	0	0	0	0
0	0	1	0	1	0	0	0	0	0	0
0	1	0	0	0	1	0	0	0	0	0
0	1	1	0	0	0	1	0	0	0	0
1	0	0	0	0	0	0	1	0	0	0
1	0	1	0	0	0	0	0	1	0	0
1	1	0	0	0	0	0	0	0	1	0
1	1	1	0	0	0	0	0	0	0	1

A partire dalla tabella di verità, esemplificata per 3 ingressi, è possibile rilevare i seguenti **legami algebrici ingresso – uscita**:

$$y_0 = \overline{x_1}x_0$$

$$y_1 = \overline{x_1}x_0$$

$$y_2 = x_1\overline{x_0}$$

$$y_3 = x_1x_0$$

Ricordando che, **sia in logica a rapporto che in logica CMOS, si è in grado di realizzare porte NAND e NOR con un solo livello logico** e che queste due costituiscono un insieme funzionalmente completo, **per realizzare un decoder su silicio si può applicare la Legge di De Morgan e circuitare le seguenti equazioni**:

$$y_0 = \overline{x_1 + x_0}$$

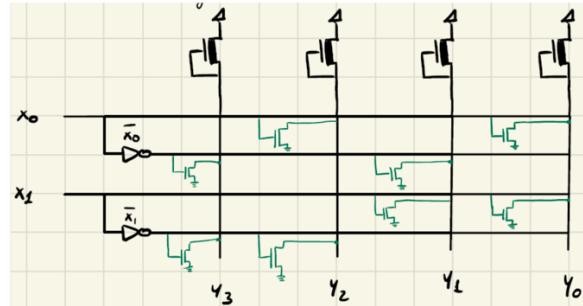
$$y_1 = \overline{x_1 + \overline{x_0}}$$

$$y_2 = \overline{\overline{x_1} + x_0}$$

$$y_3 = \overline{\overline{x_1} + \overline{x_0}}$$

Essendo presenti sia in forma affermativa che negativa, **sul silicio è necessario dotarsi di un numero di invertitori pari al numero di ingressi**, in aggiunta alle porte NOR necessarie per codificare questo

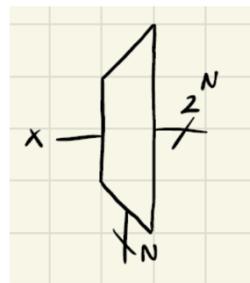
comportamento. Volendo **rappresentare graficamente un eventuale circuitazione digitale di questo decoder**, si può fare riferimento alla figura seguente:



In questa schematizzazione matriciale è possibile apprezzare la disposizione degli ingressi (con le rispettive negazioni) sulle righe e le uscite sulle colonne (che non cortocircuitano le righe su cui sono posti gli ingressi). Per una questione di ottimizzazione della figura, il circuito logico è stato realizzato in logica a rapporto (si notino gli NMOS – D in cima ad ogni colonna), consapevoli però che esiste un corrispettivo in logica CMOS.

Sebbene il passaggio da equazione algebrica di NOT e AND ad equazione di NOT, NOR e NAND possa apparire piuttosto controllintuitiva, nella realtà pratica è di fondamentale utilità, dal momento in cui la realizzazione di un circuito digitale su un solo livello logico permette di ridurre gli stadi in cascata e, quindi, l'effetto capacitivo equivalente, il quale porta inevitabilmente all'aumento del tempo di propagazione (che con questo accorgimento "complicativo" è contenuto). Infine, si noti come l'intero circuito digitale sia stato realizzato senza l'impiego di alcun bipolo che non sia un transistor (MOSFET); infatti, anche l'invertitore solamente schematizzato verrà realizzato in logica a rapporto (o in logica CMOS) e senza l'impiego di resistenze di carico (o di pull-up).

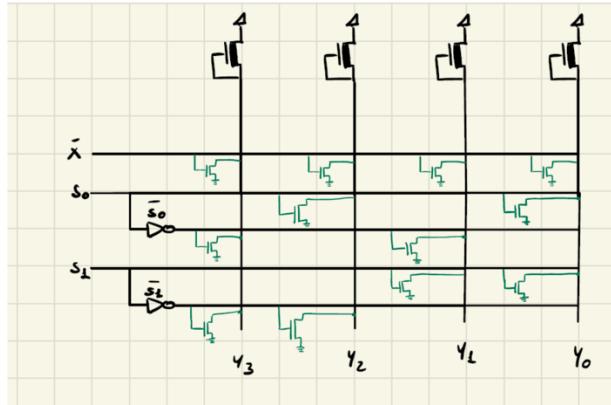
Il demultiplexer (o DEMUX) è un circuito combinatorio di instradamento, possiede un solo ingresso X che viene instradato su una delle linee 2^N linee di uscita in funzione del valore della parola di N bit di indirizzo; si può intuire che il funzionamento del DEMUX dipende fortemente da un decoder, che traduce un indirizzo di N bit nell'apertura di uno tra 2^N percorsi. Segue che la realizzazione di un circuito digitale di questo tipo non può prescindere dalla schematizzazione del decoder, dove viene sostituito l'ingresso con i bit della parola di indirizzo (S_1, \dots, S_N) e messo l'ingresso X del DEMUX in comunicazione con tutti i rami di uscita:



La tabella di verità che rappresenta il funzionamento di un DEMUX è la seguente:

s_1	s_0	A	B	C	D
0	0	IN	0	0	0
0	1	0	IN	0	0
1	0	0	0	IN	0
1	1	0	0	0	IN

Mentre la rappresentazione circuitale:



Essa deriva dalle seguenti espressioni algebriche:

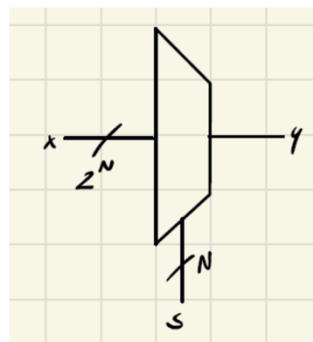
$$y_0 = \bar{s}_1 \bar{s}_0 x = \overline{s_1 + s_0 + \bar{x}}$$

$$y_1 = \bar{s}_1 s_0 x = \overline{s_1 + \bar{s}_0 + \bar{x}}$$

$$y_2 = s_1 \bar{s}_0 x = \overline{\bar{s}_1 + s_0 + \bar{x}}$$

$$y = s_1 s_0 x = \overline{\bar{s}_1 + \bar{s}_0 + \bar{x}}$$

In realtà, il **DEMUX** viene progettato come dispositivo duale al **MUX** (o **multiplexer**), un circuito di instradamento che, con 2^N ingressi, un'uscita e N bit di controllo permette di determinare da quale “strada” (tra le 2^N) prelevare l’uscita:



La tabella di verità di un MUX ad 4 bit viene così a formarsi:

$s_1 s_0$	y
0 0	x_0
0 1	x_1
1 0	x_2
1 1	x_3

Se lo si volesse astrarre un po' di più, **il comportamento di un multiplexer corrisponderebbe a quello di uno switch – case**: in funzione delle 2^N combinazioni possibili del mio dato ($s_0 s_1$), eseguo una determinata operazione piuttosto che un'altra (x_i). In altri termini:

```
switch ( $s_0 s_1$ ) {
    case 00:
        return  $x_1$ 
    case 01:
        return  $x_2$ 
    ...
}
```

Semplificando il MUX a 2 bit, la legge algebrica che descrive il comportamento del circuito è la seguente:

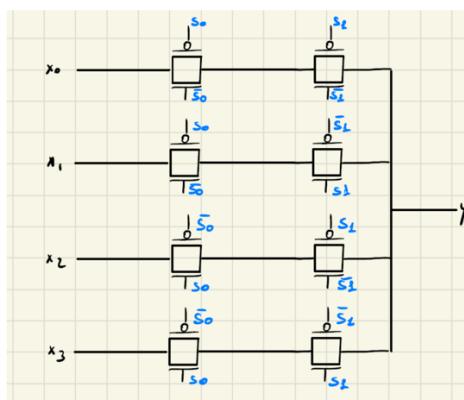
$$y = x_1 s + x_0 \bar{s}$$

Volendo trasporre il tutto **in funzione di NOT, NOR e NAND**:

$$y = (\overline{x_1} + \bar{s})(\overline{x_0} + s)$$

Tuttavia, per realizzare un MUX con questa funzione logica si impiegherebbero troppi MOSFET e, pertanto, si avrebbe un tempo di propagazione non trascurabile. Un circuito più semplice ed efficace utilizza tanti pass transistor quanti segnali da instradare; infatti, la prima funzione logica specificata permette di individuare in s il clock che abilita o meno il passaggio del segnale x_0 o x_1 e nella presenza simultanea della sua versione affermativa e negativa l'autoesclusività dei due segnali.

Ritornando a 4 bit, il MUX viene realizzato come segue:



Si sono utilizzati meno MOSFET (20) e si è avuto un minor tempo di propagazione. Si noti, che per ogni combinazione dei segnali s_0 e s_1 (con i loro negati) esiste un solo ingresso x_i che può procedere in uscita (e quindi caratterizzato da pass transistor aperti), gli altri saranno interrotti da pass transistor interdetti.

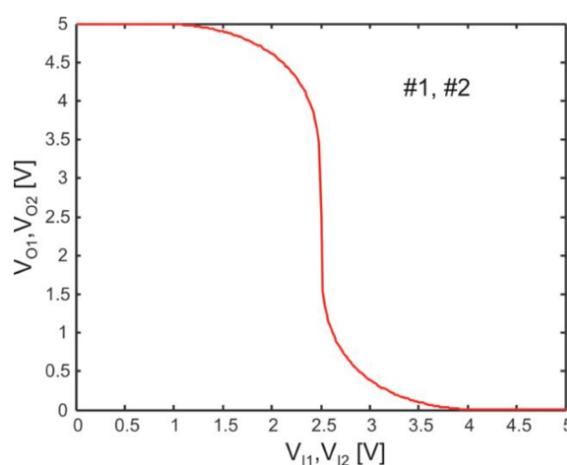
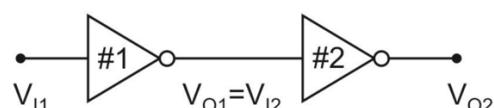
CIRCUITI SEQUENZIALI: BISTABILI, LATCH E FLIP – FLOP

È stato definito in precedenza circuito combinatorio quel particolare sistema caratterizzato da un'uscita dipendente unicamente dai valori assunti dalle variabili logiche in ingresso, senza alcuna influenza da parte di alcuna variabile di stato; di contro, un circuito sequenziale è un sistema la cui uscita dipende dalla “storia del circuito”, ovvero dai valori delle variabili logiche in ingresso (detti eventi logici) che si sono susseguiti ad istanti di tempo precedenti quello attuale. Poiché c'è una sorta di meccanismo di “memoria degli istanti precedenti”, questo tipo di circuiti viene anche detto circuito a memoria; quella parte del circuito responsabile di questa proprietà viene detta elemento di memoria e può essere categorizzato in due gruppi:

- **Elementi di memoria statici**, realizzati in modo da **autosostenere l'informazione mediante un meccanismo di rigenerazione (feedback positivo)** e, pertanto, **particolarmente robusti**, a discapito di una **maggior complessità di progettazione** e un **maggior impiego di spazio** (necessario all'autosostentimento);
- **Elementi di memoria dinamici**, realizzati tramite **l'impiego di elementi circuituali in grado di conservare temporaneamente l'informazione** (capacità, tendenzialmente), occupando **meno spazio** ma richiedendo **attenzione alla propria delicatezza**.

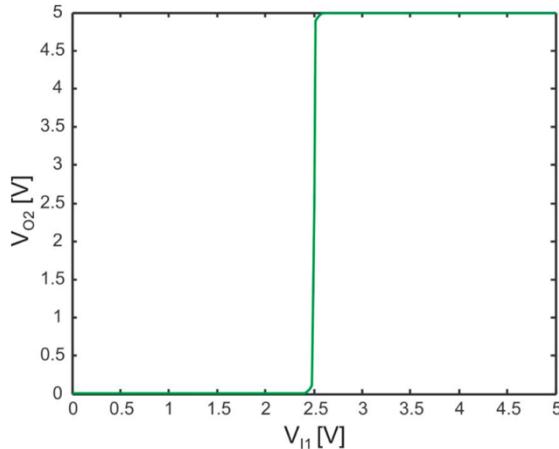
L'elemento di memoria statico di base è detto **bistabile** (o latch) e, come si può intuire dal nome, è caratterizzato dalla **presenza di due stati stabili anche in assenza di ingresso**; affermare che il circuito si trova in uno stato stabile equivale a dire che esso è in grado di permanere in quello stato anche in corrispondenza di disturbi in ingresso di modesta entità che tendono a cambiarlo.

Un primo prototipo di bistabile può essere realizzato posizionando **due invertitori in cascata** (detti anche in configurazione ad anello aperto):

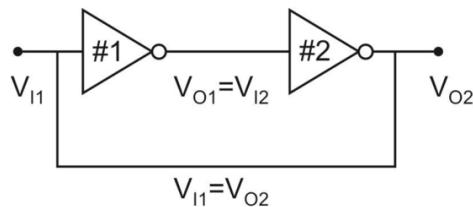


Supponendo che i due invertitori siano entrambi in logica CMOS (quindi $V_{OL} = 0$ e $V_{OH} = V_{DD}$), identici e simmetrizzati, allora le loro caratteristiche di trasferimento sono quelle specificate.

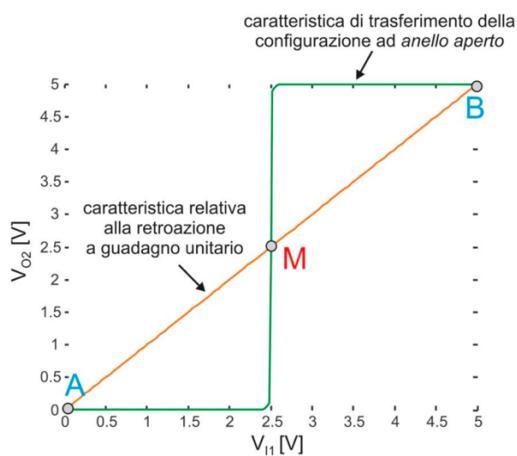
Ci si focalizzi sulla **caratteristica del dispositivo in cascata**, graficata come segue, notando che **non è una funzione logica invertente** e che la pendenza in corrispondenza di valori di ingresso intorno al punto medio di esecuzione è particolarmente elevata:



Si voglia ora analizzare la **configurazione ad anello chiuso** e dimostrare che questa, in sé e per sé, rappresenta la **schematizzazione circuitale di un bistabile**:

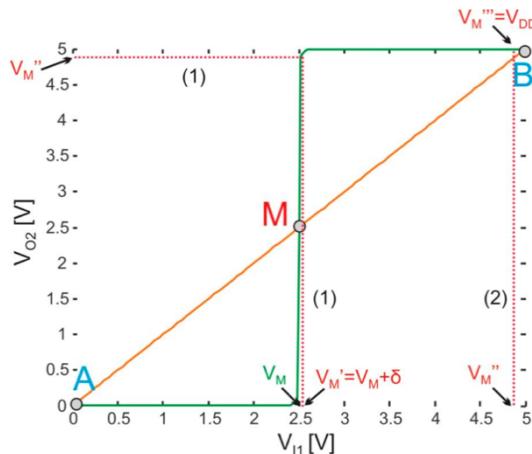


In primis, è necessario individuare i due stati di equilibrio; lo si può fare considerando la cortocircuitazione dell'ingresso e dell'uscita come un parallelo con la configurazione ad anello aperto ed andare ad osservare sul piano ausiliario le intersezioni tra la bisettrice del primo quadrante (corrispondente a $V_{I1} = V_{O2}$) e la caratteristica precedentemente graficata:



Le intersezioni rilevate sono tre, sebbene si possa dimostrare come **A e B siano stati di equilibrio asintoticamente stabile** e **M di equilibrio instabile**. Si consideri un ingresso $V_{I1} = V_M = V_{DD}/2$ e vi si sovrapponga un disturbo relativamente piccolo $\delta > 0$; l'uscita $V_{O2} = V_M''$ corrispondente a questa

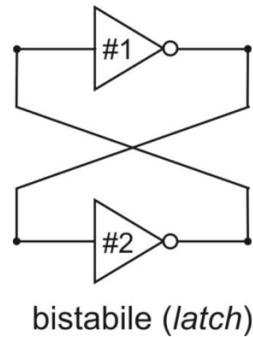
sovraposizione, $V'_M = V_M + \delta$, risulterà maggiore del punto medio in cui l'ingresso è stato supposto a causa dell'elevata pendenza della caratteristica di trasferimento ad anello aperto (1):



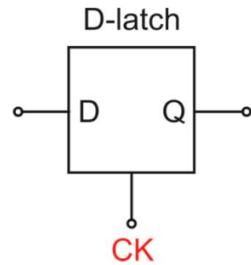
Essendoci retroazione, V''_M entrerà nuovamente in ingresso, producendo un'uscita $V'''_M > V''_M$ (2) e portando il sistema nello stato di equilibrio B. Di conseguenza, si può intuire come il punto di equilibrio M non sia stabile (e, quindi, instabile). Ovviamente, la sovrapposizione di un disturbo $\delta < 0$ al segnale $V_{I1} = V_M$ avrebbe portato il sistema verso il punto di equilibrio stabile A.

A differenza del punto M, i punti di equilibrio A e B sono asintoticamente stabili perché in corrispondenza di disturbi relativamente piccoli il sistema tenderebbe a ritornarvi; questa proprietà permette di individuare in questi due punti i due livelli logici del sistema: 0 corrispondente al punto a potenziale minore (A) e 1 a quello a potenziale maggiore (B).

In conclusione, il sistema è considerato in grado, per sua natura, di autosostenere lo stato stabile in cui si trova anche in presenza di disturbi sovrapposti ai segnali e di funzionare, pertanto, come elemento di memoria statico. Un altro modo con cui rappresentare un latch (o bistabile) elementare di questo tipo è il seguente, in cui sono evidenziati i caratteri retroattivi del sistema:



Il problema di questo sistema consiste nella sua “troppa stabilità”; infatti, non è ancora in grado di prendere in ingresso una sollecitazione tale da provocare il cambio di stato (ovvero la variazione dell'informazione autosostenuta). Una tale funzionalità applicativa viene esaustivamente implementata in un **D – latch**:

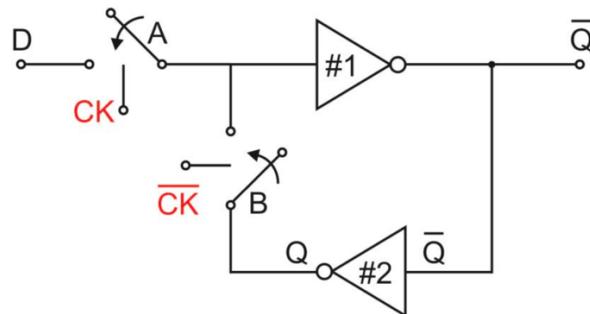


Il funzionamento è determinato interamente dal valore logico assegnato al segnale di clock (CK):

- **CK = 0, il latch si trova in fase di memorizzazione** (hold), in cui autosostiene lo stato in cui si trovava prima che il clock fosse portato a 0 indipendentemente dalle sollecitazioni di ingresso D;
- **CK = 1, il latch si trova in fase di trasparenza** (follow), in cui il latch è trasparente e l'uscita Q segue pedissequamente l'ingresso D indipendentemente dai valori di Q precedenti all'alzamento del clock a 1.

Ovviamente, le variazioni in questione si suppongono, per semplicità, **istantanee nonostante nella realtà il sistema induce un ritardo**, per quanto lieve, **non nullo**.

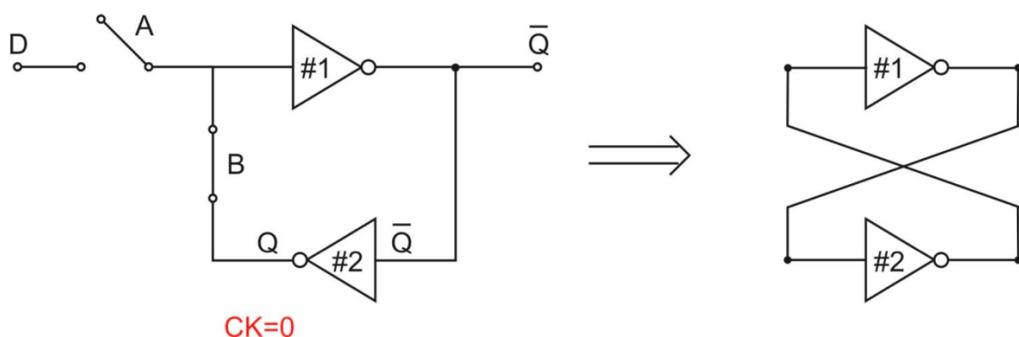
L'implementazione più compatta di un D – latch prevede l'impiego di **due pass transistor** (quindi 2 o 4 MOSFET) e di **due invertitori**, configurati come segue:



Implementazione del D-latch
tramite logica a pass transistor

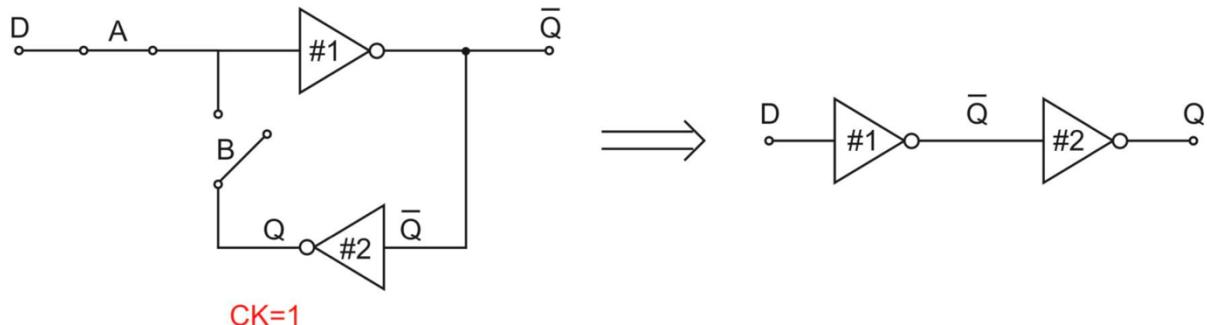
Si analizzi il comportamento di questo circuito per i due possibili valori del clock (CK):

- **CK = 0**



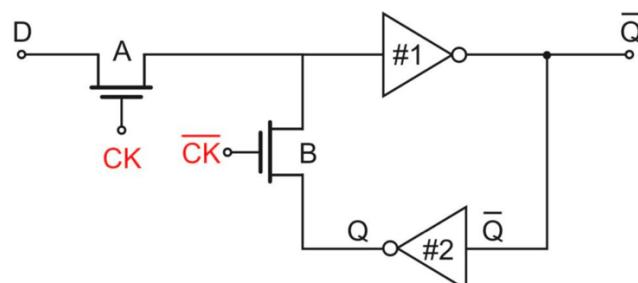
La chiusura di B e l'apertura di A portano il sistema in configurazione elementare, permettendo di assimilare il D – latch ad un semplice bistabile e di entrare in fase di hold.

- $CK = 1$



La chiusura di A connette l'ingresso D al circuito, a differenza del caso precedente, e permette a Q di adattarsi al valore di D (dopo un certo ritardo corrispondente al passaggio attraverso la coppia di invertitori); così impostato, il sistema è trasparente alle variazioni del segnale di ingresso.

Scendendo ancor di più nella pratica, in logica NMOS questo circuito diviene:

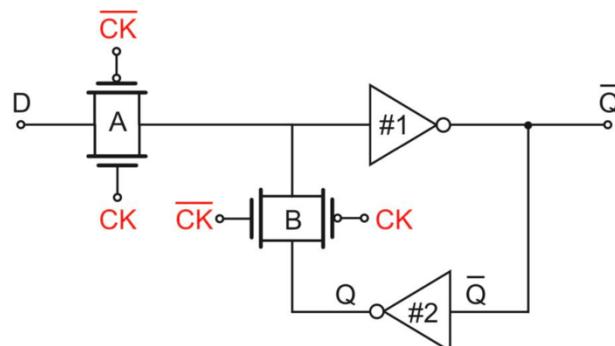


D-latch in logica a porte di trasmissione NMOS

Impiegando:

- **3 invertitori**, due visibili nella figura e uno relativo all'inversione del segnale di clock (CK);
 - **6 MOSFET**, visto che ogni invertitore richiede due MOSFET;
- **2 NMOS** montati come pass transistor;
 - Questa configurazione induce i problemi di trasmissione del 1 logico precedentemente introdotti.

Per un totale di **8 MOSFET**. Mentre in logica CMOS:

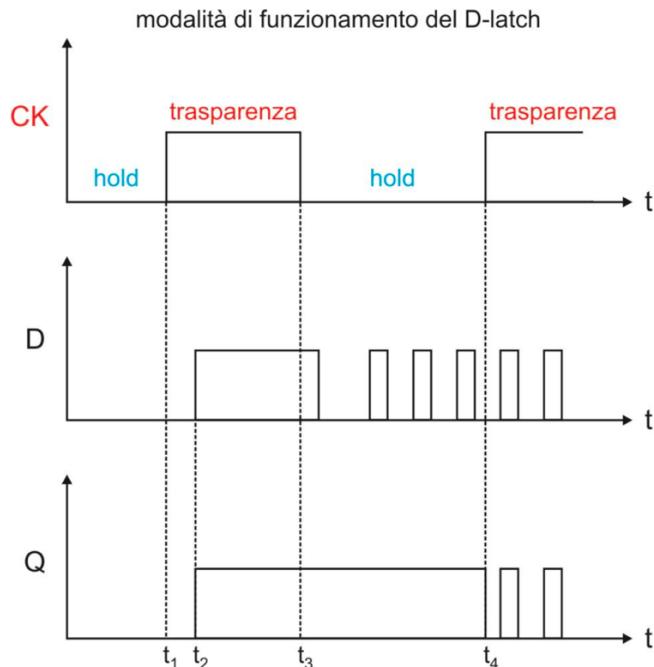


D-latch in logica a porte di trasmissione CMOS

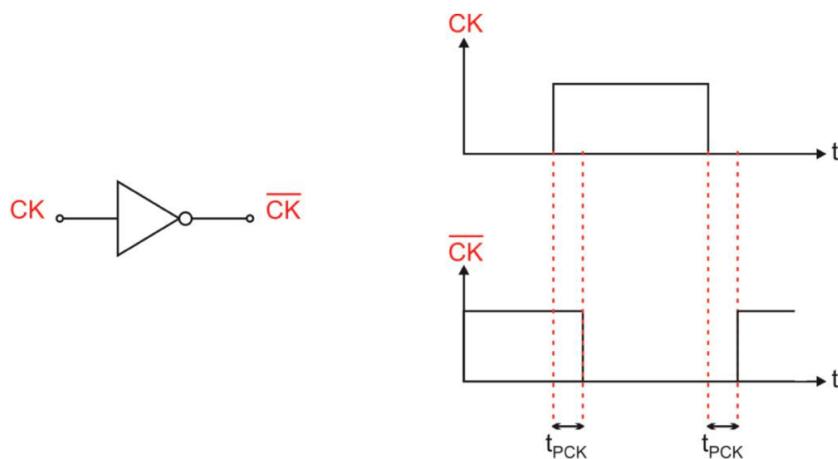
Impiegando:

- **3 invertitori**, due visibili nella figura e uno relativo all'inversione del segnale di clock (CK);
 - **6 MOSFET**, visto che ogni invertitore richiede due MOSFET;
- **2 NMOS e 2 PMOS** montati come pass transistor;
 - Questa configurazione non induce i problemi di trasmissione del 1 logico precedentemente introdotti.

Per un totale di 10 MOSFET. Il funzionamento di un D – latch può essere analizzato anche in funzione delle forme d'onda dei vari segnali coinvolti, supponendo ovviamente che le variazioni dello stato siano istantanee:



Un primo problema di questo circuito ha origine nella dicotomia tra CK e il suo negato; infatti, anche supponendo le loro variazioni istantanee, in corrispondenza di commutazioni di CK il suo negato si adatta solo dopo un tempo t_{PCK} dovuto alla porta invertente da cui è originato:

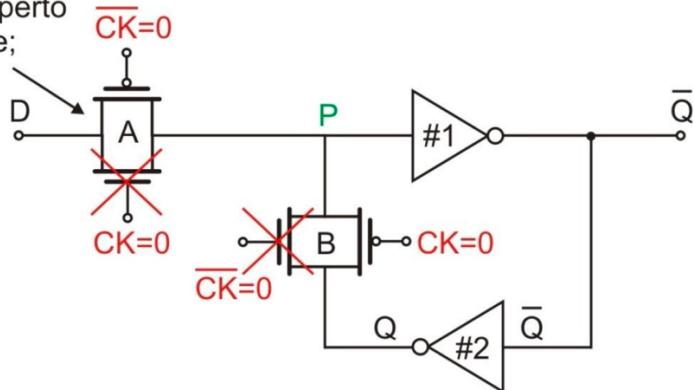


Supponendo l'invertitore simmetrizzato, in modo che $t_{PLH} = t_{PHL} = t_{PCK}$. Da questa figura si può ben osservare la **formazione di intervalli di tempo di sovrapposizione (overlap)** in cui sia CK che

il suo negato assumono lo stesso valore; queste condizioni sono particolarmente dannose per il funzionamento logico del latch e possono produrre gravi errori di elaborazione a causa della **propagazione di eventuali malfunzionamenti a stadi di collegamento successivi.**

Ad esempio, si supponga inizialmente $Q = 0$ e si voglia attivare la fase di memorizzazione commutando CK da 1 a 0. In corrispondenza degli inevitabili intervalli di sovrapposizione, gli NMOS dei pass transistor sono in interdizione e i PMOS in conduzione:

L'interruttore A dovrebbe essere aperto per favorire la memorizzazione; invece il PMOS conduce!



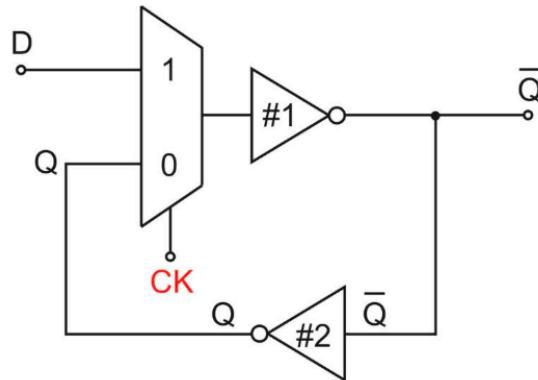
A dovrebbe essere aperto, mentre invece è chiuso e può permettere la propagazione di un eventuale cambiamento di D (da 0 a 1) a Q, andando a perdere l'informazione che si voleva memorizzare. Il problema in sé e per sé risiede nel malfunzionamento dei pass transistor, che funzionano in configurazione $CK\bar{CK} = 11 \vee 00$, che non rientrano nel dominio progettato per tale circuito; infatti, il pass transistor è progettato per funzionare da interruttore solo in condizioni di funzionamento 01 o 10 e non si comporta direttamente come tale quando invece è in configurazione 11 o 00.

Supponendo CMOS simmetrizzati gli invertitori del sistema, sia $t = 0$ l'istante di commutazione del clock (CK) da alto a basso e $t = t_D$ quello in cui il segnale di ingresso D commuta da basso ad alto. Il potenziale P inizia a salire, l'uscita del primo invertitore (\bar{Q}) a scendere e quella del secondo invertitore (Q) a salire. Nel frattempo, il PMOS della porta di trasmissione A sostiene sempre meno l'aumento di potenziale P a causa del fatto che il negato di CK sta aumentando; tuttavia, se questo aumento è molto lento (per t_{PCK} elevati), il PMOS si spegnerà quando l'uscita V_Q del secondo invertitore sarà maggiore di $V_{DD}/2$. Pertanto, il latch tende da solo a sostenere il dato sbagliato $Q = 1$, il che significa che D = 1 ha avuto il tempo di propagarsi nella fase di pseudo-trasparenza.

Al fine di evitare il cambio di stato, può essere buona norma contenere il t_{PCK} entro il t_{PHL} del primo invertitore, in modo da evitare che la tensione corrispondente al negato di Q si porti ad un valore minore di $V_{DD}/2$. In tal modo, si è sicuri che il circuito non è in grado di "sostenere" la propagazione di D = 1 per un tempo sufficiente. Nella pratica, le condizioni appena enunciate possono essere ottenute evitando che lo stesso circuito invertente del clock (CK) piloti molti pass transistor per l'ingresso CK negato; infatti, in tal caso, la capacità parassita delle porte in cascata aumenterebbe a tal punto da rendere t_{PCK} non più basso come desiderato. È chiaro come la soluzione più conveniente sia quella che implementa più invertitori locali, invece che uno globale, associati ad ogni latch individuale, aumentando però il numero di MOSFET e lo spazio richiesti dal circuito ma rendendo irrilevanti gli intervalli in cui CK = \bar{CK} .

Si vuole concludere la trattazione sul D – latch menzionando una piccola osservazione; si consideri un D – latch implementato in logica a porte di trasmissione come mostrato in precedenza.

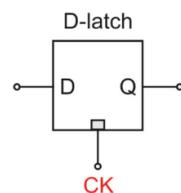
Ricordando che il multiplexer è un blocco che, a seconda della combinazione di valori assunti da una parola di indirizzo di N bit, instrada sull'uscita uno tra 2^N dati di ingresso; pertanto, il D – latch prototipizzato può essere schematizzato anche come segue:



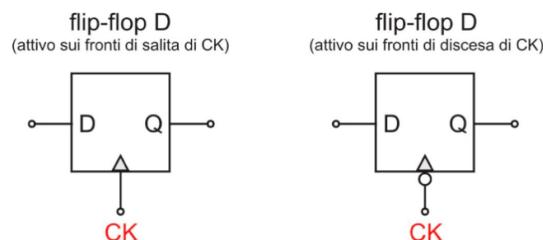
Infatti:

- Se $CK = 0$, viene instradato sull'uscita del MUX (cioè sull'ingresso del primo invertitore) il dato corrispondente al bit di selezione 0, ovvero Q , ottenendo così il bistabile elementare ed entrando in **fase di hold**;
- Se $CK = 1$, viene instradato sull'uscita del MUX (cioè sull'ingresso del primo invertitore) il dato corrispondente al bit di selezione 1, ovvero D , che viene riportato su Q dopo un ritardo legato alla cascata di invertitori ed entrando in **fase di follow**.

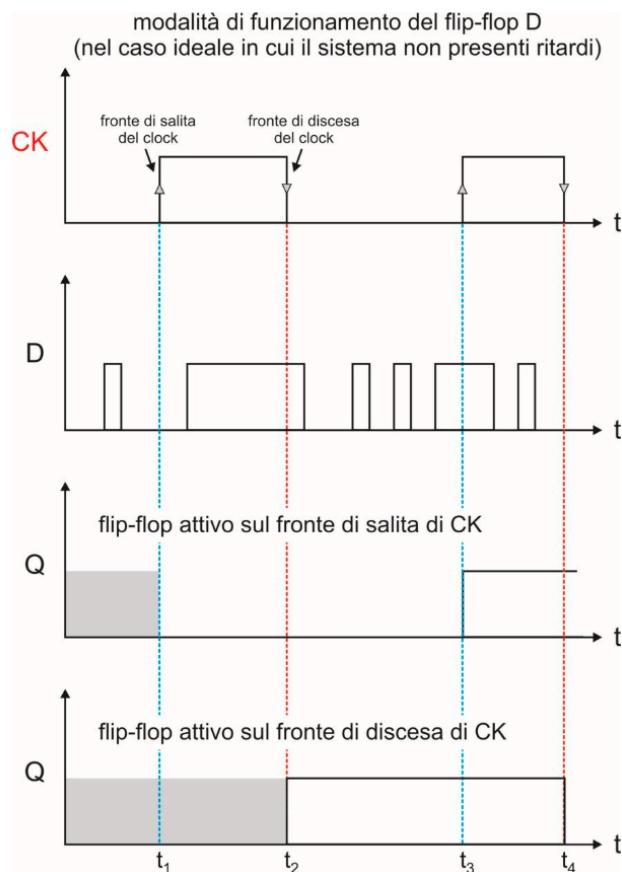
Il D – latch (così come altri circuiti simili, tipo il SR – latch non approfondito in questa sede) sono realizzati in modo tale che il loro funzionamento dipenda dal livello logico del clock: se è 1 si è in una condizione di funzionamento, se è 0 ci si trova in un'altra. Quando un sistema segue questo comportamento è detto **funzionalmente dettato dal livello logico** e, graficamente, lo si può individuare dal **quadratino posto sopra al segnale di clock**:



Di contro, i flip – flop si trovano in una determinata condizione di funzionamento piuttosto che in un'altra sulla base del fronte di salita (o di discesa) del clock, non del livello logico del clock stesso; sistemi come questi sono detti **funzionalmente dipendenti dai fronti di salita/discesa** e, come sarà a breve più chiaro, non sono mai trasparenti come i latch. Graficamente, questi sistemi sono individuati da un **triangolino sul segnale di clock**, accompagnato dal classico cerchietto per distinguere il funzionamento attivo sui fronti di salita da quello sui fronti di discesa:

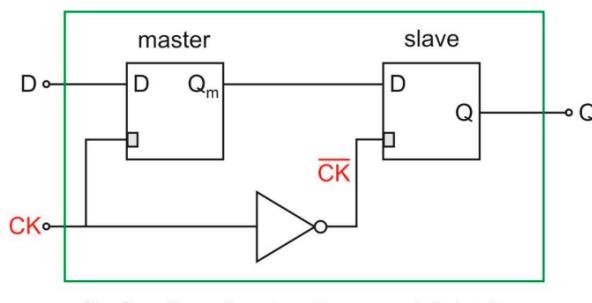


Supponendo il flip – flop D sensibile ai fronti di salita del clock, in corrispondenza della commutazione $0 \rightarrow 1$ il dato D viene “catturato” e “collocato” su Q in una seconda fase in cui D non è più “visto” dal circuito.



Gli andamenti temporali dei segnali coinvolti qui mostrati possono chiarire il funzionamento di un flip – flop D. Nel caso di **flip – flop attivo sui fronti di salita**, ciò che accade a Q prima dell’istante di tempo t_1 è **indifferente** (analogamente per l’istante t_2 nel caso di flip – flop attivo sui fronti di discesa).

Sul fronte di salita per $t = t_1$, l’ingresso D è basso e tale dovrà essere anche Q (si è in fase di trasparenza), dopodiché i due segnali divengono indipendenti, con Q che permane basso (si è in fase di memorizzazione); di contro, sul fronte di salita $t = t_3$, l’ingresso D è alto e il valore di Q va a **seguire** (si è in fase di trasparenza), dopodiché i due segnali divengono indipendenti, con Q che permane alto (si è in fase di memorizzazione). Si può fare un **discorso analogo anche per il flip – flop attivo su fronti di discesa**.



flip-flop D realizzato attraverso 2 D-latch
in configurazione master-slave

Un modo piuttosto semplice per realizzare un flip – flop D è in configurazione master – slave, composta da due D – latch in cascata, uno attivo su un livello del clock e uno attivo sull’altro.

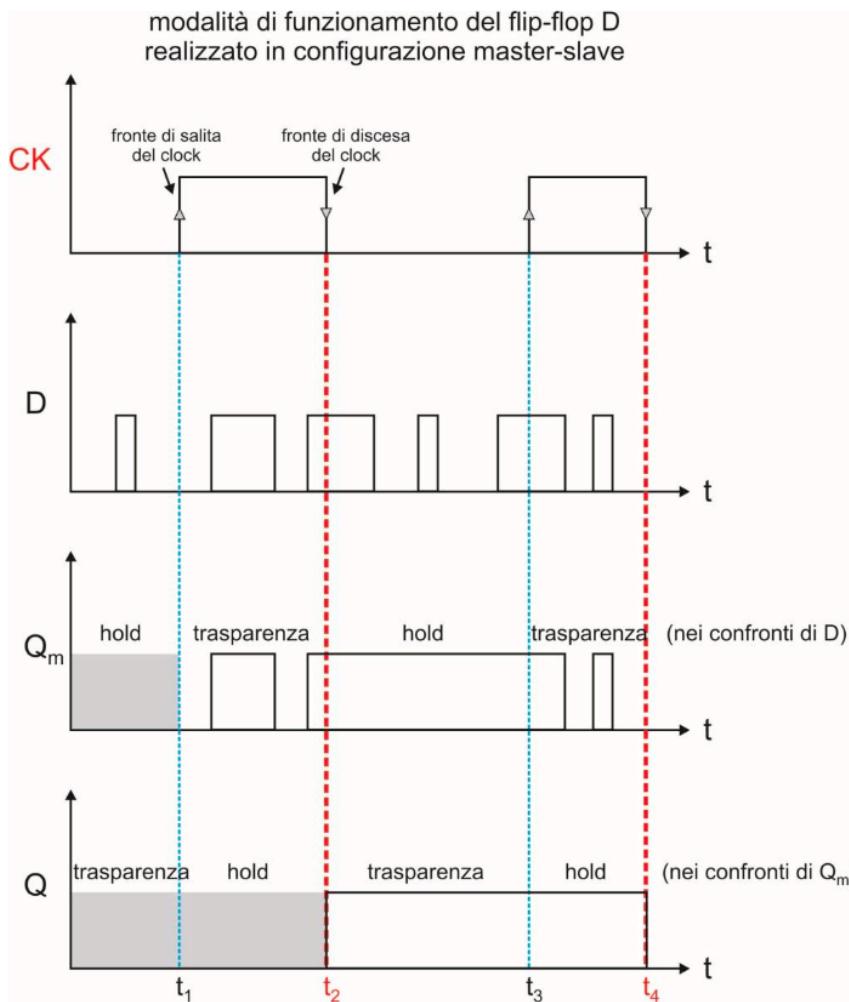
Affinché questa configurazione circuitale si comporti da flip – flop D bisogna dimostrare che:

- **Il circuito non è mai trasparente;**
- **Il circuito è sensibile ai fronti di salita (o di discesa) piuttosto che al livello logico;**

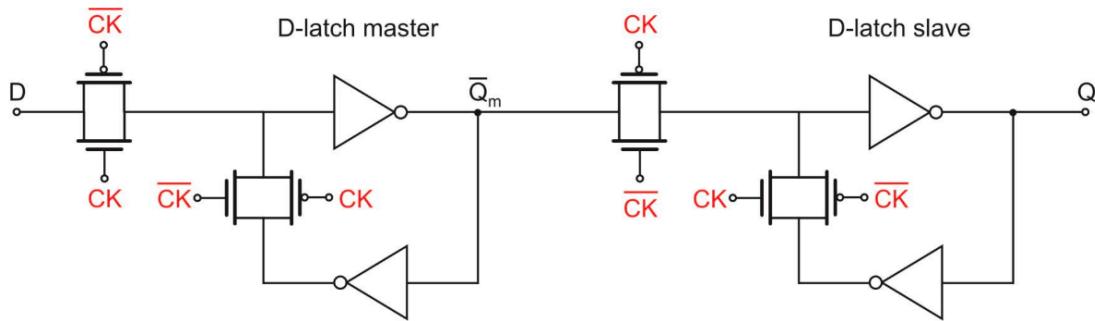
Se **CK = 1** si ha il D – latch master in trasparenza ma il slave è in memorizzazione; ciò equivale a dire che Q_m segue con un certo ritardo le variazioni di D ma non fa altrettanto Q. Quando **CK = 0** si ha il D – latch master in memorizzazione, bloccando l’uscita Q_m al valore che aveva finché CK era alto, mentre lo slave diviene trasparente e riporta su Q il valore che prima della commutazione era su D.

Il circuito non è mai trasparente perché tra master e slave c’è sempre uno stadio in memorizzazione che non propaga l’uscita attraverso sé stesso, facendo in modo che D e Q siano collegati direttamente solo in un istante (quello di commutazione di CK). Inoltre, da quanto appena mostrato, se la commutazione del clock (CK) avviene nel verso $1 \rightarrow 0$ il master si porta da trasparenza a memorizzazione, congelando il valore della propria uscita Q_m a quello di ingresso D, e lo slave da memorizzazione a trasparenza, portando all’uscita Q il valore appena memorizzato dal master, ovvero l’ingresso D; il circuito è funzionalmente sensibile al fronte di discesa.

Il tutto può essere apprezzato anche graficamente:



Dal punto di vista circuituale:



Realizzazione di un flip-flop D master-slave in logica a porte di trasmissione CMOS

Si noti che **l'ingresso del D – latch slave è il negato del segnale Q_m** visto nei diagrammi temporali mostrati in precedenza (è ottenuto da D a meno di un inversione), mentre **il segnale Q in uscita dallo stesso latch coincide con il segnale Q mostrato nei suddetti diagrammi** (è ottenuto da D attraverso una doppia inversione).

Il numero complessivo di MOSFET necessari per un'implementazione di questo tipo è dato dalla somma di quelli usati nei 5 invertitori (10) e quelli relativi alle 4 porte di trasmissione CMOS (8), per un **totale di 18 MOSFET**.

CIRCUITI SEQUENZIALI: MEMORIE NON VOLATILI

Nella realizzazione di una qualsiasi macchina computazionale **la realizzazione dei circuiti di memoria riveste un ruolo centrale al corretto funzionamento di una qualsiasi elaborazione da eseguire**; infatti, è possibile dimostrare come **ogni circuito digitale sequenziale può essere schematizzato dall'insieme di un circuito combinatorio e da un elemento circuitale in grado di memorizzare lo stato del sistema**.

La memorizzazione di un bit di informazione può essere effettuata, in una prima analisi, **in due maniere differenti**: può essere implementato un **circuito in grado di mantenere uno tra due stati stabili quando opportunamente forzato** (ad esempio un latch o un flip – flop) **o definendo la memorizzazione di un livello logico binario come la presenza o assenza di carica immagazzinata in un condensatore**.

Le memorie possono essere classificate sulla base di due criteri, per tempo di accesso:

- **SAM (Sequential Access Memory)**, sono tali che il tempo di accesso ad un informazione è dipendente dalla posizione della stessa:
 - Sono **tecnologie ormai in disuso** e più diffuse agli albori dell'informatica, ad esempio con la macchina a nastro di Turing;
- **RAM (Random Access Memory)**, sono tali che il tempo di accesso ad un informazione non dipende in alcun modo dalla posizione della stessa:
 - È la tecnologia con cui è **realizzata la maggior parte delle memorie al giorno d'oggi**;
 - Si pensi che con l'aumento delle dimensioni delle memorie, costruire una SAM equivale a realizzare un dispositivo che ci mette molto più tempo ad accedere al dato rispetto a quello che l'utilizzatore impiega per usare il dato stesso.

Quindi, si può intuire come **commercialmente la differenziazione tra memoria RAM e memoria di archiviazione sia dettata unicamente dall'abitudine**, dal momento in cui **entrambe saranno realizzate come memorie RAM**. Mentre per operatività:

- **RWM (Read and Write Memory)**, sono memorie in cui è possibile sia scrivere che leggere;
- **ROM (Read Only Memory)**, sono memorie in cui è possibile solo l'operazione di lettura:
 - Sono memorie **poco usate in ambito commerciale** e relegate a ruoli di primaria importanza ma non accessibili all'utilizzatore medio
- **WOM (Write Only Memory)**, sono memorie in cui è possibile solo l'operazione di scrittura:
 - Sono **piuttosto inutili**, se si scrive qualcosa da qualche parte è perché la si vuole conservare per una futura lettura.

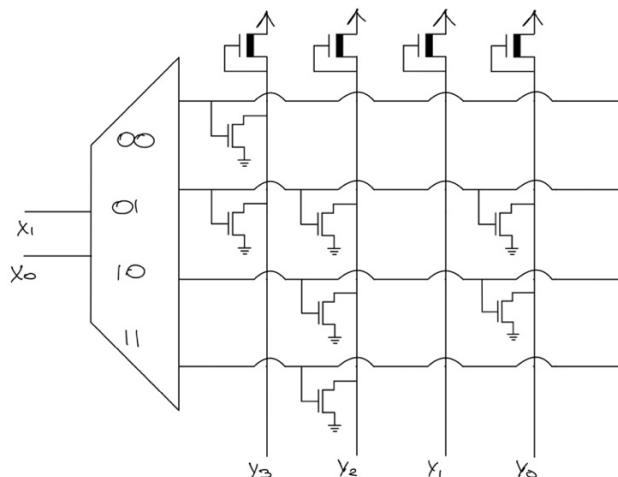
Volendo essere precisi, **le memorie con cui si ha a che fare la maggior parte delle volte sono memorie RAM RWM**; tuttavia, poiché ormai tutte le memorie sono RAM, questo acronimo verrà omesso. Volendo fare un'ultima precisazione, **per classificare le memorie sulla base della loro operatività non è sufficiente descrivere quali operazioni sono possibili ma è utile specificare anche la probabilità con cui tali operazioni sono eseguite**; infatti, **le memorie RWM che sono commercialmente usate nei dispositivi computazionali sono progettate in modo che sia molto più probabile l'accesso ad un'informazione che la sua scrittura** (o meglio, ci sono un numero di scritture ingegneristicamente nullo rispetto al numero di letture). Quindi, **pur essendo memorie RWM, si comportano perlopiù da memorie ROM** (si dice siano memorie ROM quasi ovunque).

Concludendo questa introduzione, **si affronteranno in questa sede le memorie RAM ROM**, cioè ad accesso casuale e di sola lettura. **Poiché ingegneristicamente si legge più che scrive**, il circuito che descriverà il comportamento di questa memoria sarà **combinatorio** (un comportamento sequenziale si ha solo quando c'è da "ricordare" ciò che si scrive).

Procedendo per gradi di astrazione sempre più bassi, **una memoria è prima di tutto un dispositivo di instradamento che prende in ingresso un indirizzo di N bit e restituisce una tra 2^N informazioni di M bit**; pertanto, si può intuire che **la capacità di una memoria è individuata dalla quantità**:

$$M \cdot 2^N$$

Una memoria ad 2 byte può essere costruita in modo che le parole di indirizzo siano composte da 2 bit e i dati da 4 bit ($M \cdot 2^N = 4 \cdot 2^2 = 16 = 2 \cdot 8 = 2\text{byte}$). Scendendo ancora di più nel concreto, la memoria può essere vista come un decoder a cui è collegata a cascata una maglia circuitale di invertitori che portano in uscita una delle due tensioni di soglia logiche:

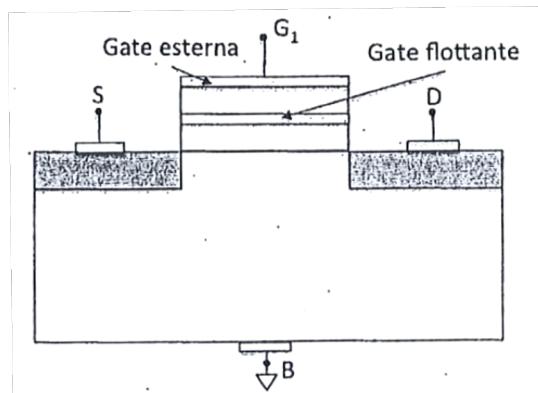


Laddove il circuito di ingresso è collegato a quello di uscita mediante il NMOS, la rete di pull – down porta sull’uscita V_{OL} , codificando il valore logico 0, mentre se l’uscita non è collegata all’ingresso vi cade sopra tutta la tensione di alimentazione V_{DD} , codificando il valore logico 1. Ad esempio, richiedendo l’informazione in posizione $x_1x_0 = 10$ verrà restituito sulle linee di uscita $y_3y_2y_1y_0 = 1010$; quindi, la presenza o assenza del MOSFET codifica la presenza, rispettivamente, di uno 0 o di un 1.

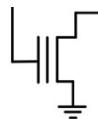
Quella appena raffigurata è una vera e propria memoria ROM, visto che agli invertitori di ogni singola cella non è possibile applicare alcun ingresso che cambi lo stato interno della memoria. Come anticipato, memorie di questo tipo non sono del tutto in disuso, possono ancora oggi essere necessarie delle applicazioni in cui i software scritti devono essere particolarmente robusti e immutabili (quindi la memorizzazione non deve poter subire alterazioni) come nel settore della difesa o della sicurezza.

Per abilitare il circuito alla scrittura è possibile implementare un particolare tipo di MOSFET che si comporta come NMOS sotto determinate condizioni (specificando una rete di pull – down e codificando il valore logico 0) ma che viene inibito sotto altre (non specificando alcuna rete di pull – down e codificando il valore logico 1). Questo tipo di MOSFET è detto FAMOS (Floating gate Avalanche MOS) e sfrutta un meccanismo di shifting della tensione di soglia per poter pilotare la conduzione del dispositivo (e quindi la rete di pull – down) sulla base del valore della V_{GS} applicata.

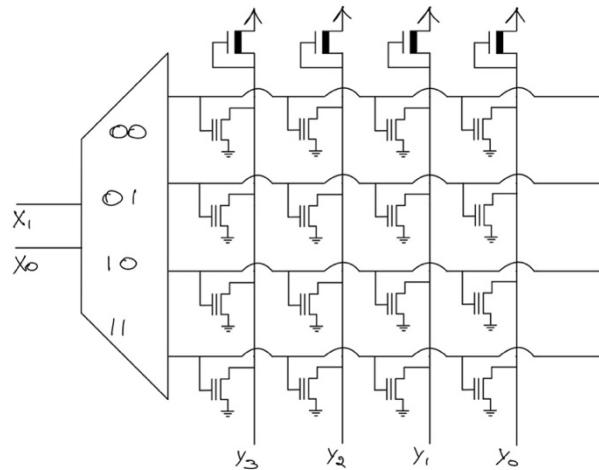
La realizzazione di un FAMOS, come è possibile intuire leggendo l’acronimo, passa per l’implementazione di una seconda Gate, chiamata floating (o flottante) perché non è connessa né al terminale di Gate, né a quello di Source o Drain e né al semiconduttore, è isolata da due strati di ossido così come in figura:



Ed è rappresentato dal seguente simbolo circuitale:



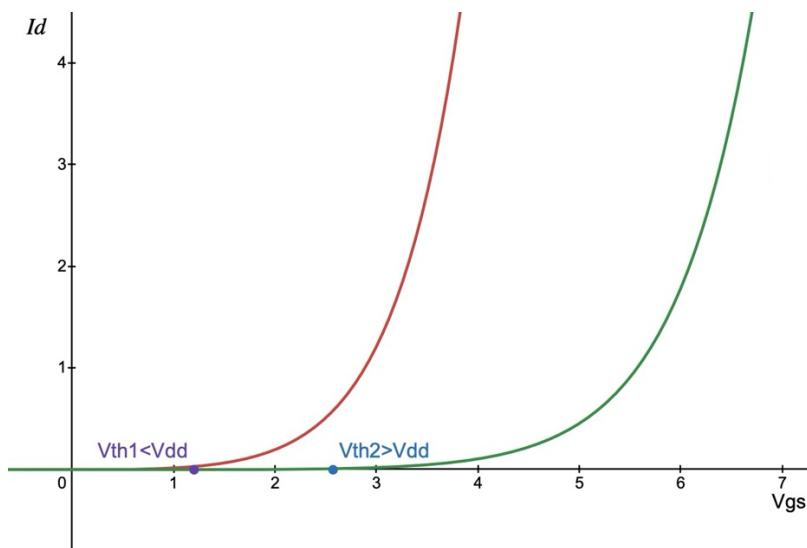
Non molto differente dai MOSFET finora analizzati. Per implementare una memoria ROM che permetta la scrittura, così come precedentemente specificato, utilizzando i FAMOS si ricorre al seguente schema circuitale:



Volendo approfondire il funzionamento di un FAMOS, si tenga in considerazione il fatto che un qualsiasi MOSFET è “acceso” quando $V_{GS} > V_{th}$; di conseguenza, per poter inibire il dispositivo (non si può agire su V_{GS} perché controlla la rete di pull – down) si deve agire aumentando V_{th} , ovvero incrementando i portatori nella floating Gate.

Infatti, per portare il MOSFET in conduzione, si deve creare il canale di conduzione nel semiconduttore a ridosso dell’ossido accumulandovi cariche negative e creando uno squilibrio per il quale sulla Gate si trovano solo cariche positive (creando così la capacità). Se prima di creare il canale conduttivo ci fossero già delle cariche negative sulla Gate (in questo caso flottante), servirebbe più tensione V_{GS} per la creazione del canale di conduzione (quindi una V_{th} maggiore) perché bisognerebbe prima bilanciare tali cariche dalla Gate per lasciar spazio a quelle positive e poi accumulare sui due terminali di questo falso capacitatore cariche negative e positive e creare, infine, il canale conduttivo. Ovviamente si ricordi che per carica positiva si intende assenza di carica negativa in un metallo.

Per una data V_{DS} , si può visualizzare così l’incremento della V_{GS} :



Pertanto, modulando all’occorrenza il numero di elettroni nella floating Gate in modo tale che V_{th} sia maggiore o minore di V_{DD} , si discrimina la possibilità di escludere o no la rete di pull – down e di rappresentare uno dei due livelli logici. I metodi con cui possono essere accumulati elettroni sulla Gate flottante sono due:

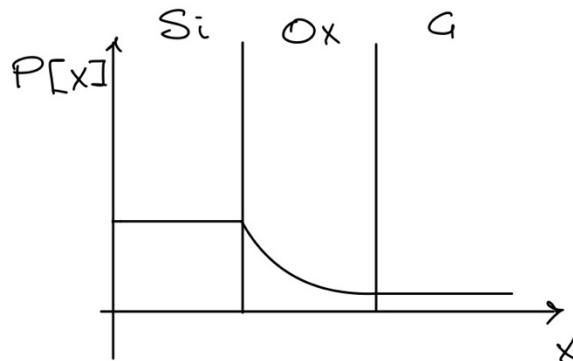
- Hot state electron (metodo di fisica classica, energia cinetica)

Inizialmente, il canale è creato applicando una V_{GS} molto elevata (che può anche duplicare V_{DD} tramite circuiti booster che, occasionalmente, erogano tensioni maggiori di quella d'alimentazione); dopodiché, è applicata una V_{DS} anch'essa molto elevata, in modo tale che gli elettroni siano accelerati. L'accelerazione in questione è diretta sia verso il Source che verso la Gate (essendo sia V_{GS} che V_{DS} elevate) e, se l'energia cinetica associata è abbastanza elevata, gli elettroni sono in grado di rompere l'ossido e giungere nella floating Gate (si comportano come veri e propri proiettili).

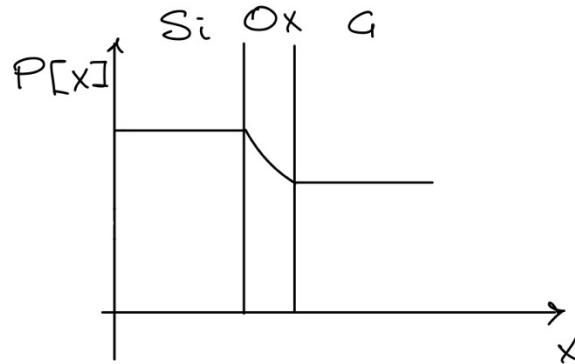
Il problema di questo tipo di memorie è l'eliminazione del dato; infatti, una volta che gli elettroni sono nella floating Gate, non sono più in grado di muoversi perché non c'è abbastanza spazio per produrre un'accelerazione (e quindi un'energia cinetica) tale da rompere in verso opposto l'ossido e tornare nel semiconduttore (sono considerati "fermi"). Le memorie che implementano questa tecnologia sono, pertanto, dette EPROM (Electrically Programmable ROM) ed escludono la possibilità di eliminare il dato (non Erasable); in realtà, esistono dei dispositivi separati, detti Eraser, che sottopongono la memoria a raggi UV per qualche secondo, in modo che l'ossido diventi temporaneamente conduttivo, facendo tornare gli elettroni nel semiconduttore. Questo processo di eliminazione, però, non è selettivo e non può essere utilizzato commercialmente per ovvi motivi; inoltre, per queste memorie esiste un numero di programmazioni massime da poter effettuare (circa 100 – 200), visto che ad ogni programmazione si bombarda l'ossido con numerosi proiettili che lo danneggiano irreparabilmente.

- Effetto tunnel (metodo di fisica quantistica, energia potenziale)

L'effetto tunnel permette di spostare un elettrone oltre l'ossido senza muoverlo; il principio su cui si basa tutto è la duplicità della natura di un elettrone, sia ondulatoria che particellare, per la quale è possibile associarvi una funzione d'onda (o funzione di Shrodinger) che descrive la probabilità che una qualsiasi particella si trovi in una determinata posizione. Inizialmente, l'elettrone nel semiconduttore è caratterizzato da una probabilità non nulla (ma bassa) di trovarsi a ridosso dell'ossido; tale probabilità subisce un decadimento nella regione dell'ossido e produce una probabilità non nulla ma ingegneristicamente inutilizzabile di trovare l'elettrone nella Gate:



Si può intuire come, per poter essere sicuri di trovare l'elettrone oltre l'ossido, si debba aumentare la probabilità in questione, in modo che il decadimento produca poi una probabilità di trovare l'elettrone nella Gate sempre bassa ma almeno utilizzabile:



Ciò può essere **ottenuto in due modi**: avvicinando la gate al semiconduttore, quindi **diminuendo lo spessore dell'ossido in modo da ridurre il decadimento**, o **aumentando l'energia potenziale dell'elettrone**, quindi **aumentando di molto la tensione V_{GS}** (anche a più del doppio della tensione di alimentazione). Con questi accorgimenti, la **probabilità di trovare l'elettrone nella Gate aumenta fino ad 1 su 1 milione**; si può pensare che tale probabilità sia ancora inutilizzabile ma tenendo in considerazione l'elevato numero di elettroni nel semiconduttore (ben oltre il miliardo) si può intuire come non sia illogico pensare di trovare dopo un “breve” periodo di tempo un numero utile di elettroni oltre l’ossido.

Il vantaggio di questa tecnologia rispetto alla controparte classica consiste nella possibilità di programmare ed eliminare le informazioni contenute nella memoria **quante volte si vuole e dove si vuole; infatti, l'**eliminazione non solo è possibile** (si ragiona allo stesso modo ma in verso opposto, quindi si deve applicare una $V_{SG} = -V_{GS}$ molto elevata) **ma è anche selettiva** (si possono eliminare le informazioni in determinati settori o in determinate celle).** Le memorie realizzate sulla base dell’effetto tunnel sono dette **memorie FLASH** e sono **quelle industrialmente prevalenti al giorno d’oggi** (sebbene abbiano ricevuto nel tempo delle migliori prestazioni).

Si vuole concludere con una piccola precisazione; **ci si chiede perché la possibilità di scrittura non sia possibile senza la floating Gate**, con un normale MOSFET. Il perché risiede nella **corrente di Gate, necessariamente nulla per garantire il corretto funzionamento del MOSFET**, e, se non si inserisce un conduttore isolato dal terminale di Gate (nel FAMOS è la floating Gate stessa) **in cui accumulare elettroni per aumentare la V_{th}** , si portano cariche negative verso un terminale che non è isolato e che poi va ad **indurre una corrente laddove non dovrebbe esserci**.

