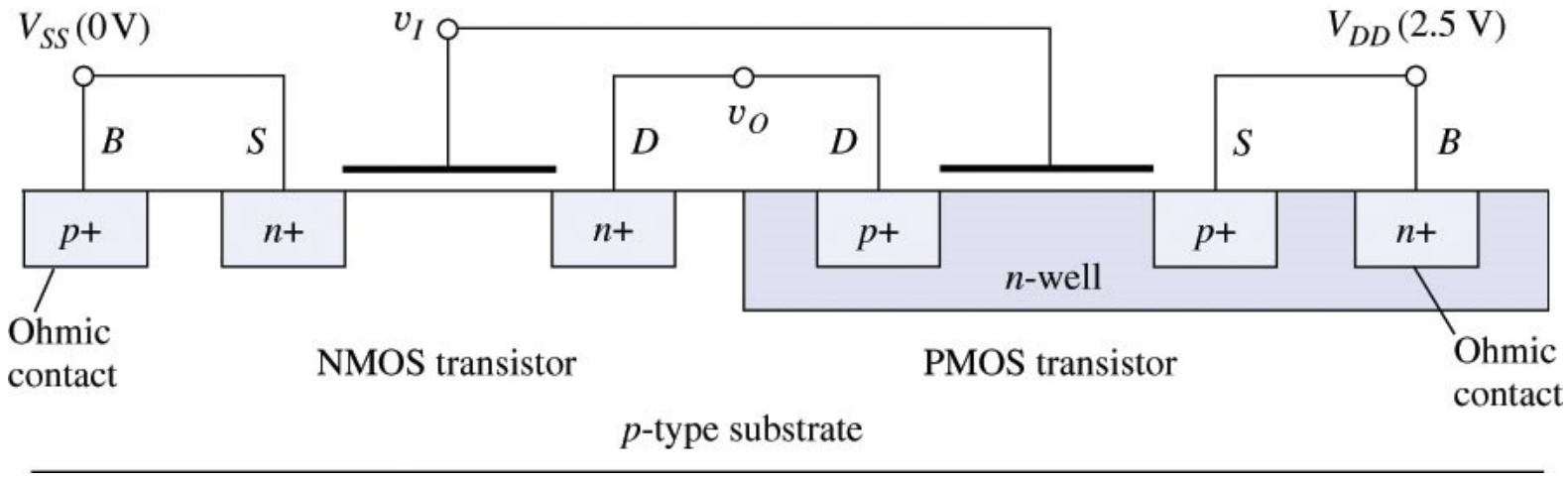
Invertitore CMOS

Circuiti CMOS

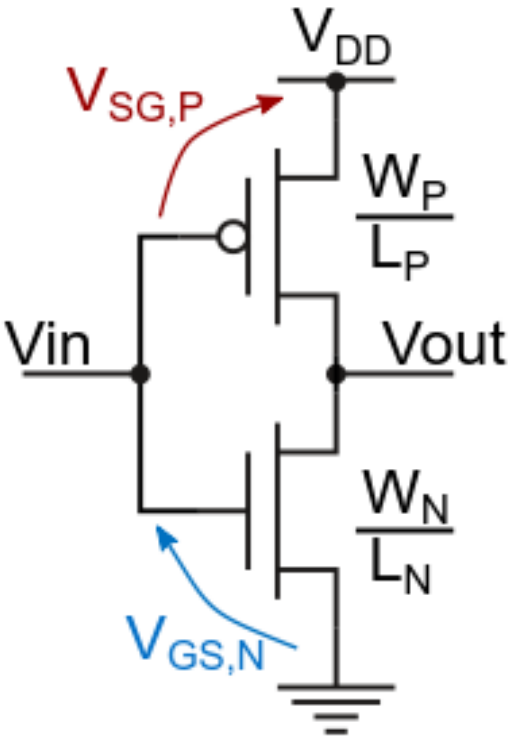
- I circuiti CMOS (Complementary MOS) utilizzano un egual numero di dispositivi NMOS e PMOS per realizzare le porte logiche
- I circuiti CMOS furono concepiti nel 1963 by Wanlass and Sah, ma sono divenuti di uso comune solo a partire dagli anni '80 quando divennero indispensabili per ridurre la dissipazione di potenza dei microprocessori NMOS
- **I circuiti CMOS sono quelli di gran lunga più adoperati negli odierni circuiti VLSI.**

Tecnologia CMOS

- È possibile realizzare su di uno stesso wafer sia dispositivi NMOS sia PMOS partendo da un substrato di tipo P in cui si realizzeranno gli NMOS.
- All'interno del substrato **si crea una ampia zona di tipo N**. In questa regione N (denominata N-well) vengono realizzati i dispositivi PMOS.
- Si può procedere in maniera duale, partendo da un substrato N e realizzando una tasca P in cui allocare i dispositivi PMOS.



CMOS Inverter

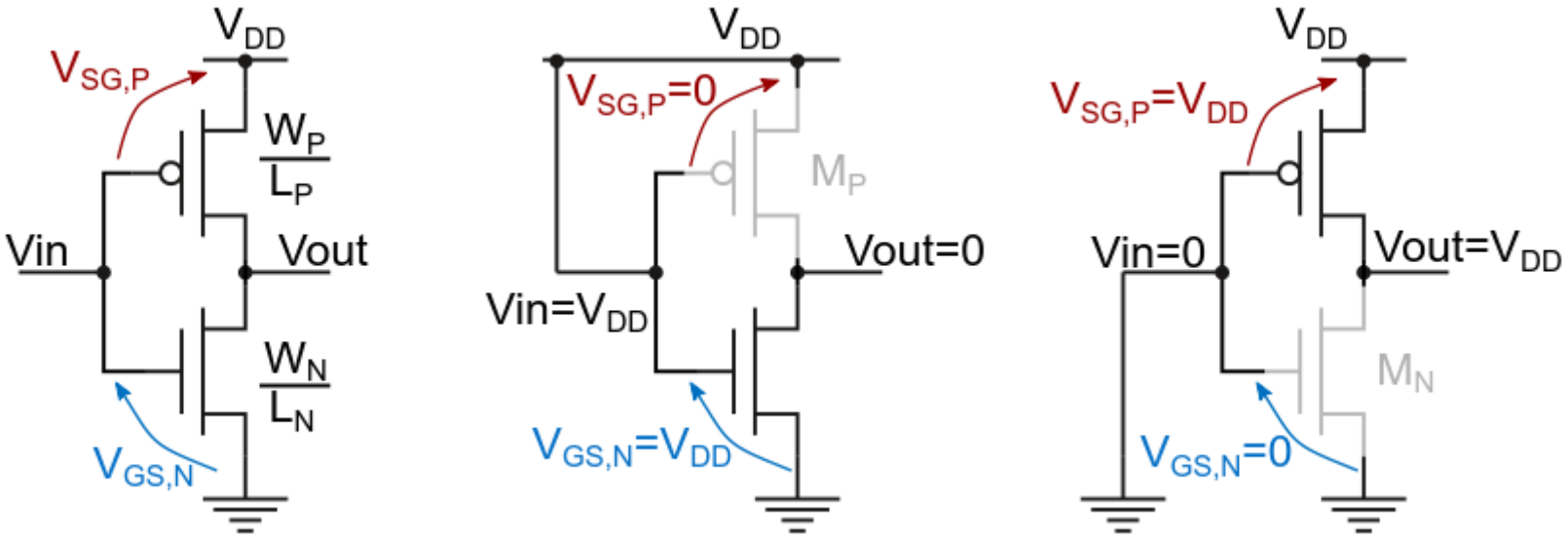


Nell'invertitore CMOS, entrambi i dispositivi sono collegati dal segnale di ingresso.

Si noti che il segnale di comando del NMOS è: $V_{GS,n}=V_{DD}$

mentre il segnale di comando del PMOS è: $V_{SG,p}=V_{DD}-Vin$

CMOS Inverter



Quando il segnale di ingresso è alto, NMOS è in conduzione mentre il PMOS è spento.

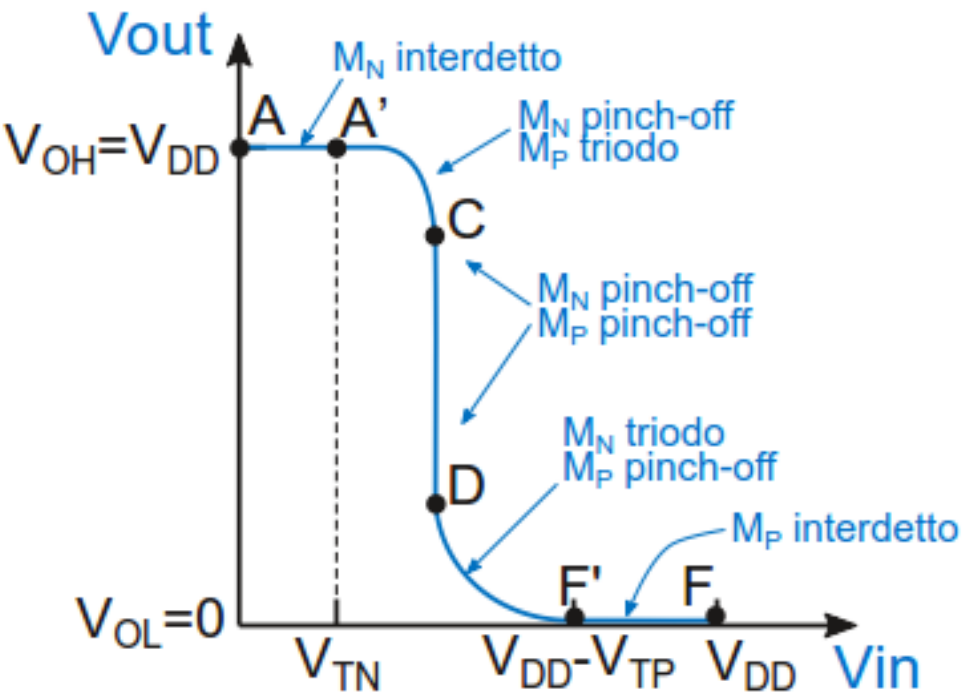
Viceversa, per ingresso basso il PMOS conduce mentre il NMOS è OFF

Funzionamento dell'invertitore CMOS

- Quando si applica un ingresso alto (V_{DD}), il PMOS è spento mentre il dispositivo NMOS è in conduzione e porta l'uscita a 0.

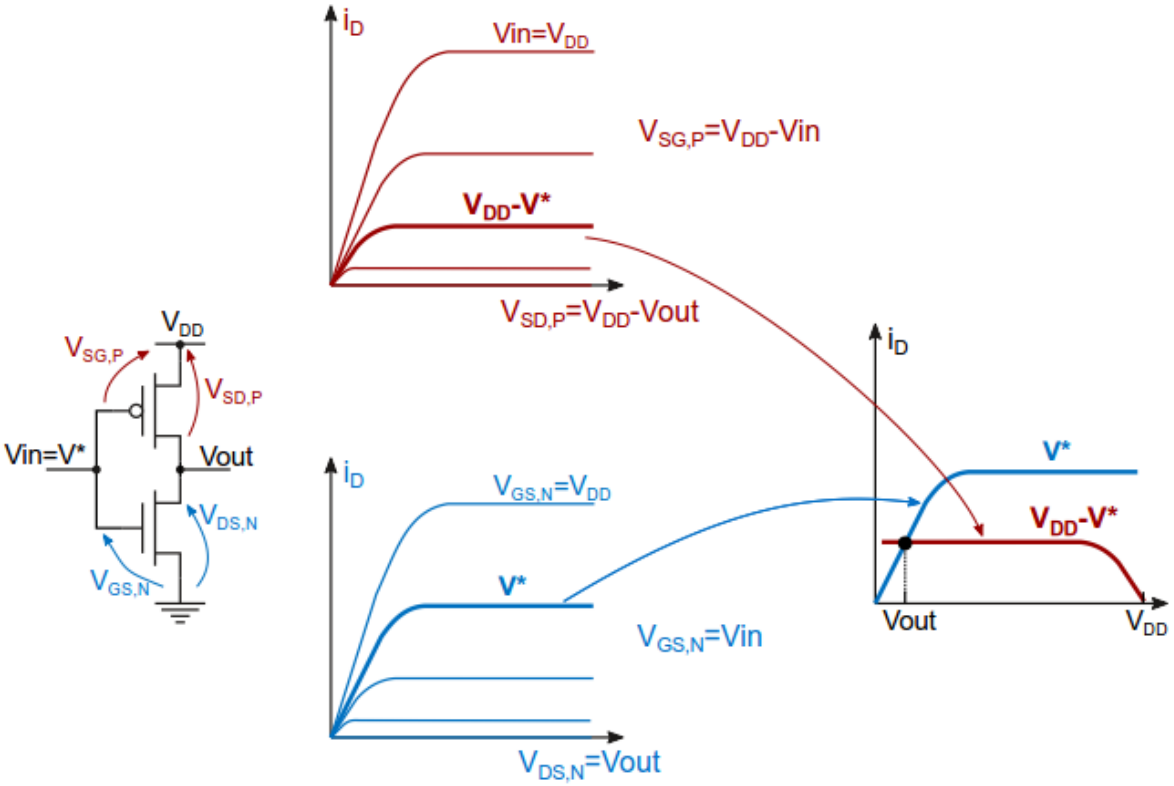
- Viceversa, quando si applica un ingresso basso, il dispositivo NMOS è spento mentre il PMOS è in conduzione e porta l'uscita a V_{DD}
- Possiamo dunque avere: $V_{OH} = V_{DD}$ e $V_{OL} = 0V$; inoltre la corrente assorbita in condizioni stazionarie è nulla e pertanto non vi è dissipazione di potenza statica

Caratteristica di trasferimento



La figura mostra una tipica caratteristica di trasferimento per un inverter CMOS.
Si possono individuare 5 regioni di funzionamento.
I livelli logici sono $V_{OH}=V_{DD}$ e $V_{OL}=0V$

Caratteristica di trasferimento



Per ottenere la caratteristica di trasferimento ci poniamo nel piano I_D,Vout.
Per ogni valore della tensione di ingresso determiniamo la corrispondente caratteristica del NMOS e del PMOS.
L'intersezione delle due curve ci darà la tensione di uscita.
Da notare che, per come è collegato il dispositivo, le caratteristiche del PMOS originano dal punto di coordinate V_{DD},0

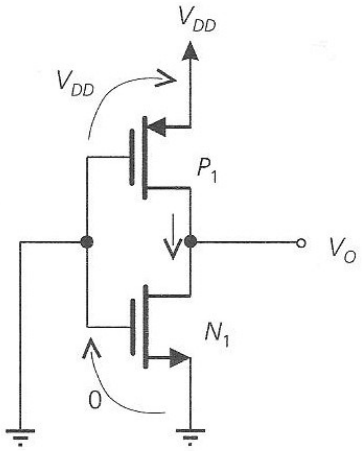
Caratteristica di trasferimento

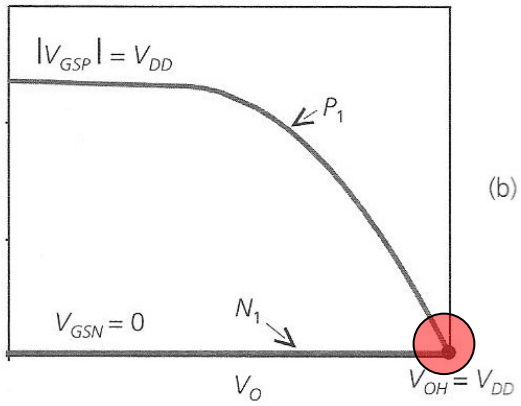
Nel seguito, per semplicità, assumeremo che le tensioni di soglia di NMOS e PMOS siano uguali in modulo, ed indicheremo entrambe le grandezze con V_T:

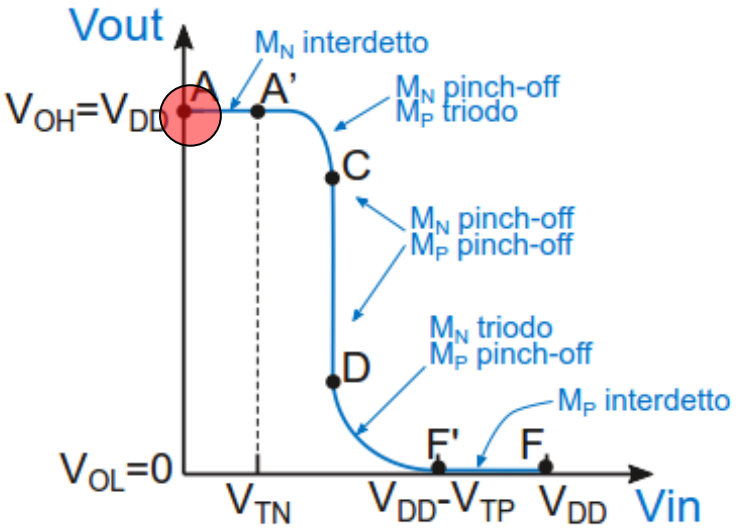
$$V_{TN} = |V_{TP}| = V_T$$

Caratteristica di trasferimento

Consideriamo la condizione Vi=0.
Le caratteristiche dei due dispositivi mostrano che:M_N=off;
M_P:on Vo=V_{DD}
Nulla cambia se Vi aumenta ma rimane inferiore a V_T

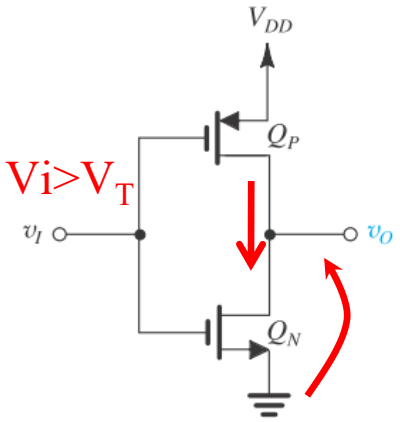


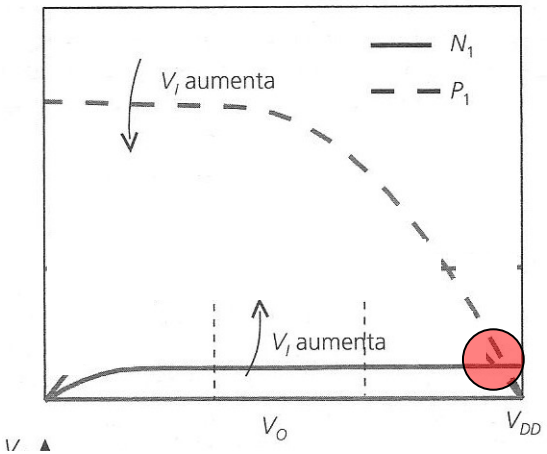


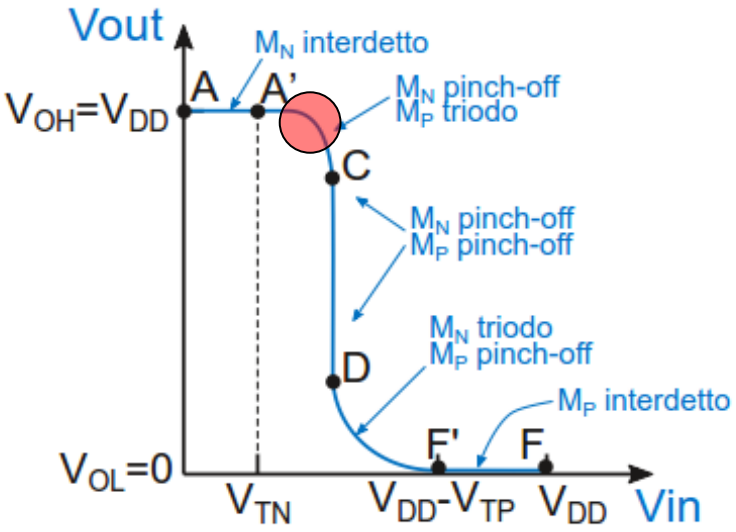


Caratteristica di trasferimento

Aumentiamo Vi oltre V_T.
M_N entra in conduzione e Vo comincia a diminuire.
In questo caso:
M_N=pinch-off; MP:triodo

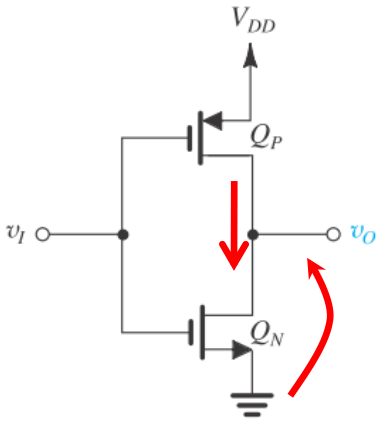


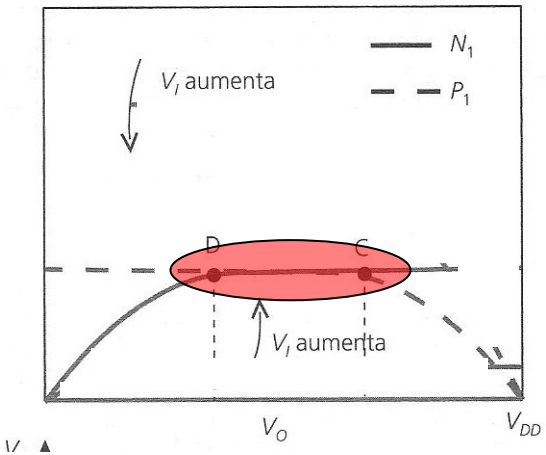


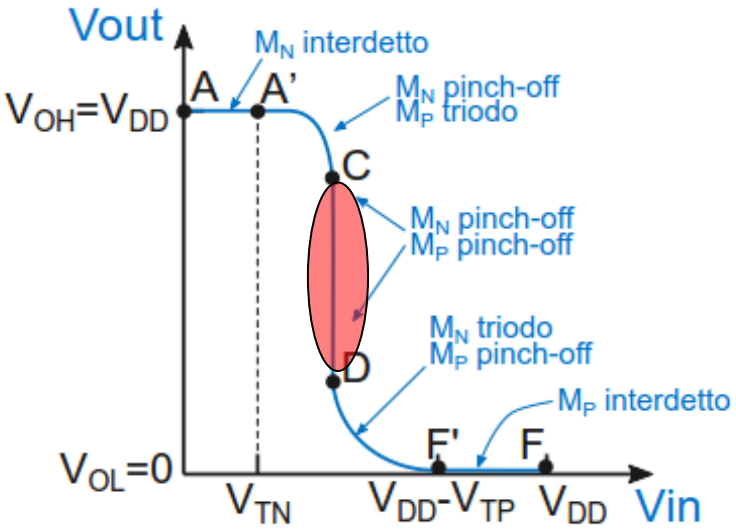


Caratteristica di trasferimento

Aumentiamo ulteriormente Vi.
Si raggiunge una condizione in cui entrambi i dispositivi sono in pinch-off: c'è un tratto pressoché verticale nella caratteristica di trasferimento.

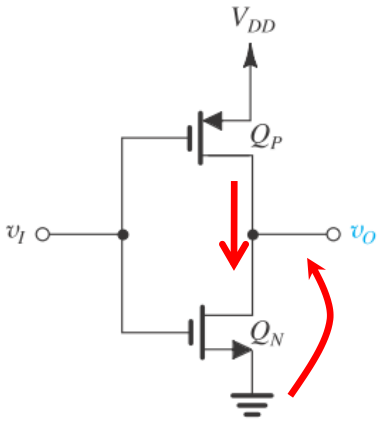


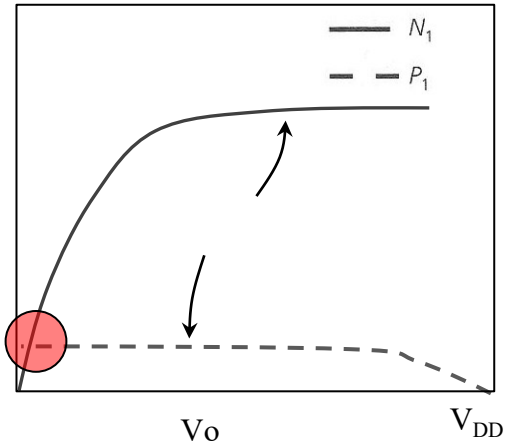


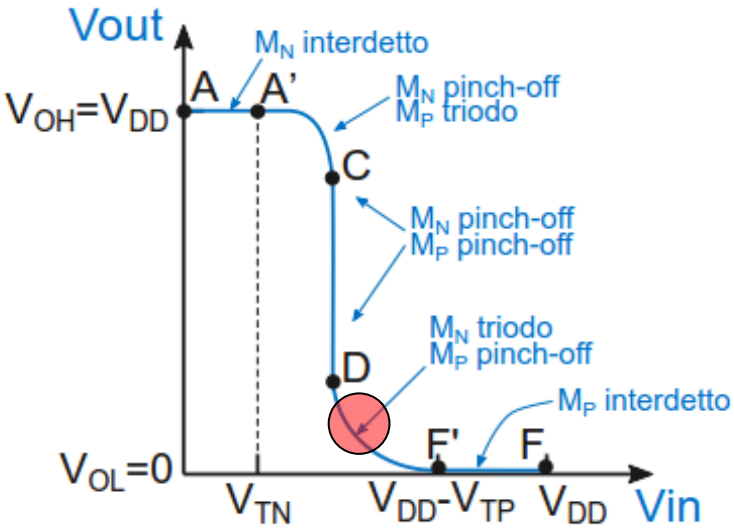


Caratteristica di trasferimento

Un ulteriore aumento di Vi porta M_N in triodo, mentre M_P resta in pinch-off

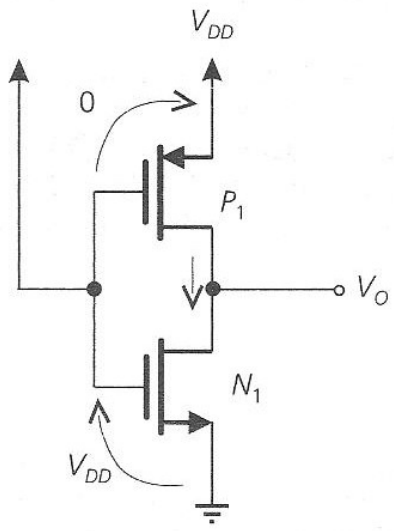


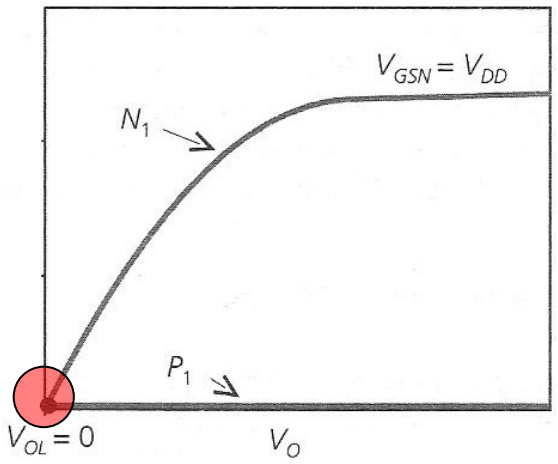


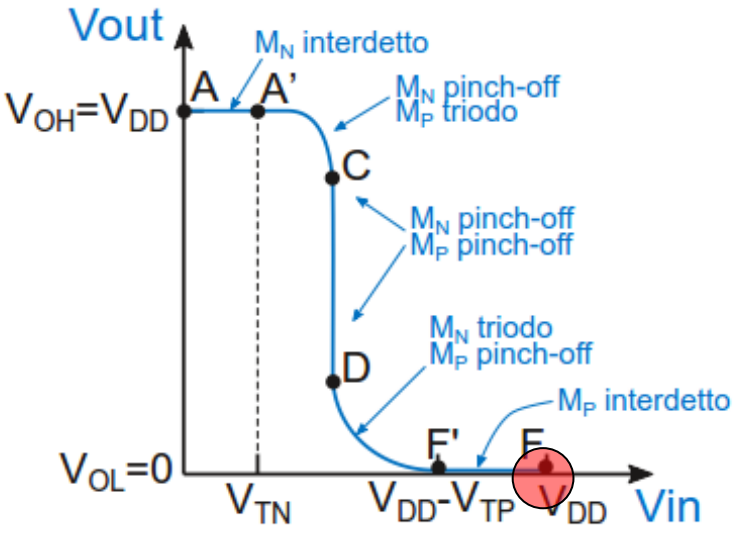


Caratteristica di trasferimento

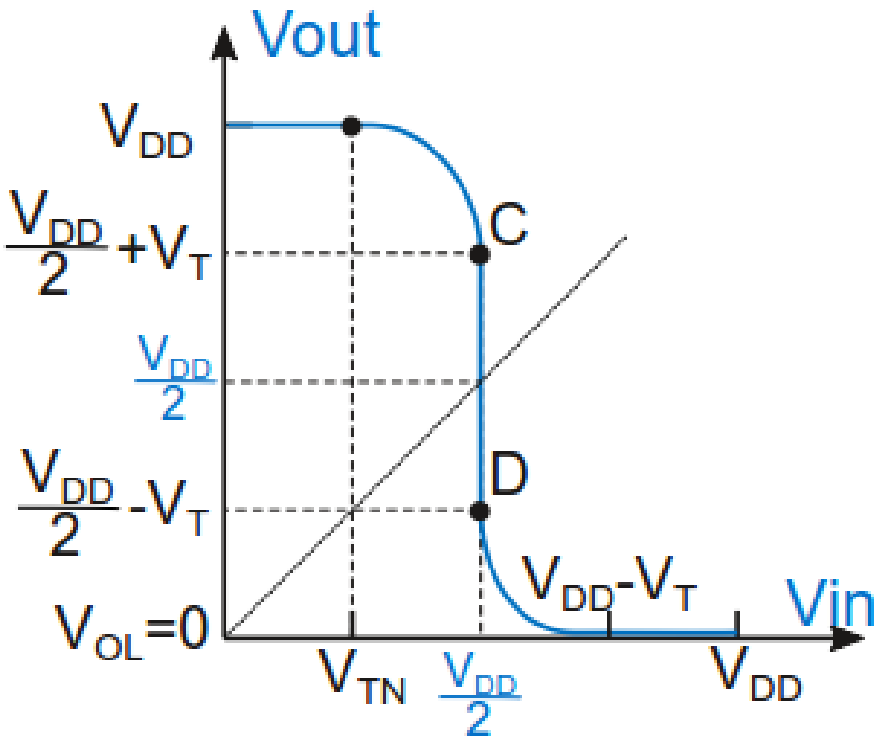
Consideriamo la condizione $Vi=V_{DD}$.
Le caratteristiche dei due dispositivi mostrano che Vo=0
M_P:off; M_N:on.
La stessa situazione si ha se $Vi>V_{DD}-V_T$







Caratteristica di trasferimento



La caratteristica di trasferimento è simmetrica quando le caratteristiche dei dispositivi sono uguali: $V_{TN}=|V_{TP}|$ e Kn=Kp

L'eguaglianza dei K si ottiene agendo su i rapporti W/L dei due MOS:

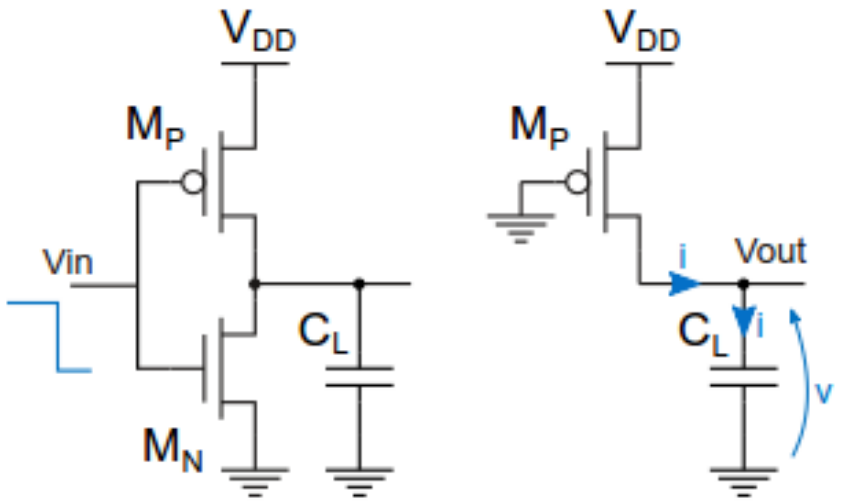
$$\left(\frac{W}{L}\right)_P = \frac{\mu_N}{\mu_P}\left(\frac{W}{L}\right)_N$$

assumeremo: $\frac{\mu_N}{\mu_P} \simeq 2.5$ e quindi per invertitore simmetrico: $\left(\frac{W}{L}\right)_P \simeq 2.5\left(\frac{W}{L}\right)_N$

Ritardo di propagazione

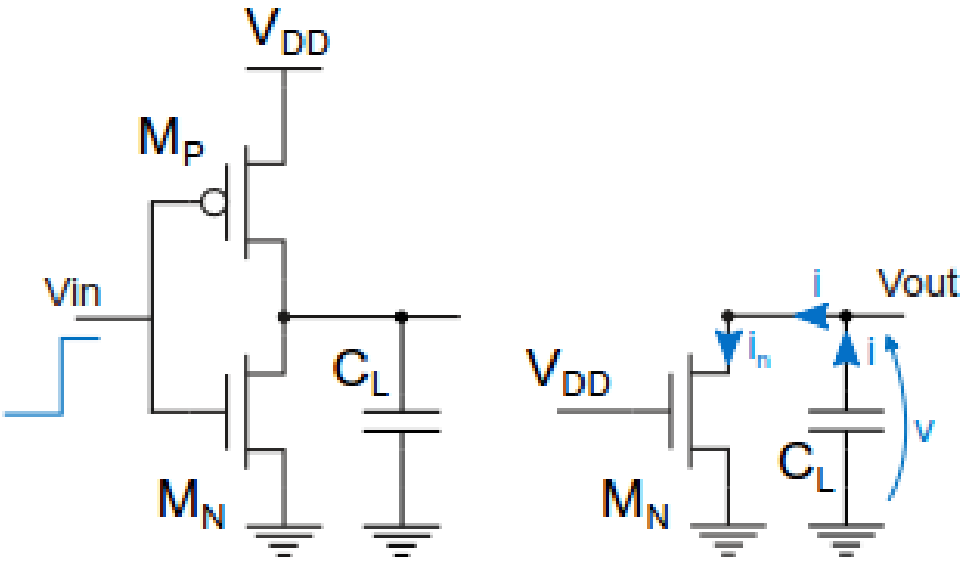
Ritardo di propagazione

Durante la transizione basso → alto dell'uscita, la capacità di uscita si carica attraverso il PMOS, mentre lo NMOS è interdetto



Ritardo di propagazione

Durante la transizione alto → basso dell'uscita abbiamo la condizione opposta: la capacità di uscita si scarica attraverso lo NMOS, mentre il PMOS è interdetto



Ritardo di propagazione

I tempi di propagazione si possono stimare assumendo che i dispositivi in conduzione siano in pinch-off durante la carica/scarica della capacità.

Ritardo di propagazione

Calcolo del tphl:

$$I \simeq \frac{K_n'}{2}\left(\frac{W}{L}\right)_N \left(V_{DD} - V_T\right)^2$$

$$I = C\frac{\Delta V}{\Delta t} \quad \text{(I è costante)}$$

$$\Delta t = C\frac{\Delta V}{I} \quad \Rightarrow \quad t_{phl} = C\frac{Vdd/2}{I}$$

$$t_{phl} = C\frac{Vdd/2}{\frac{K_n'}{2}\left(\frac{W}{L}\right)_N \left(V_{DD} - V_T\right)^2}$$

Analogamente si ottiene il t_{plh}:

$$t_{plh} = C\frac{Vdd/2}{\frac{K_p'}{2}\left(\frac{W}{L}\right)_P \left(V_{DD} - V_T\right)^2}$$

Ritardo di propagazione

Se l'invertitore è simmetrico $V_{TN}=|V_{TP}|=V_T$; Kn=Kp; i tempi di propagazione tp_{hl} e tp_{lh} sono uguali.

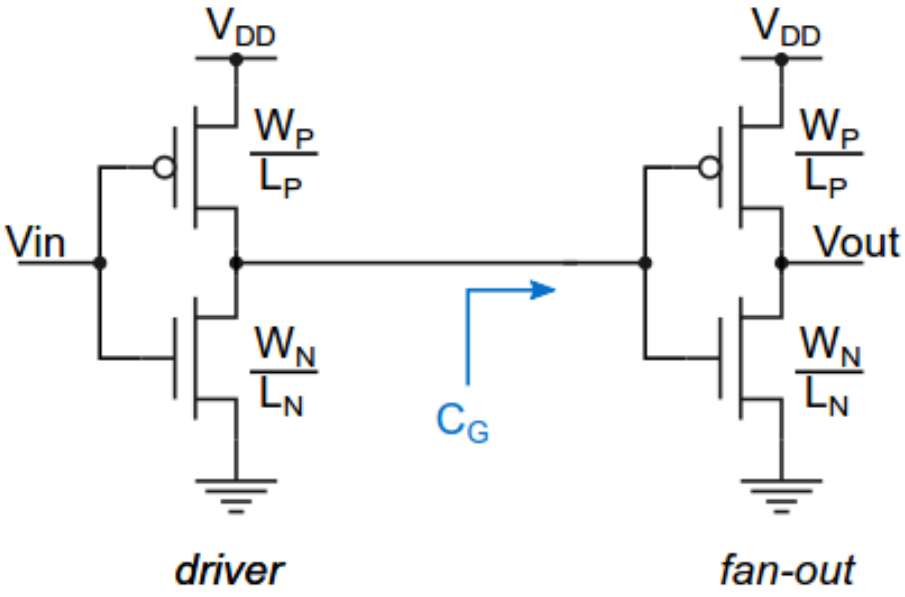
Ricordiamo che:

$$Kn = Kp = K \Leftrightarrow K_n^{'}\left(\frac{W}{L}\right)_n = K_p^{'}\left(\frac{W}{L}\right)_p \Leftrightarrow \left(\frac{W}{L}\right)_p = \frac{\mu_n}{\mu_p}\left(\frac{W}{L}\right)_n$$

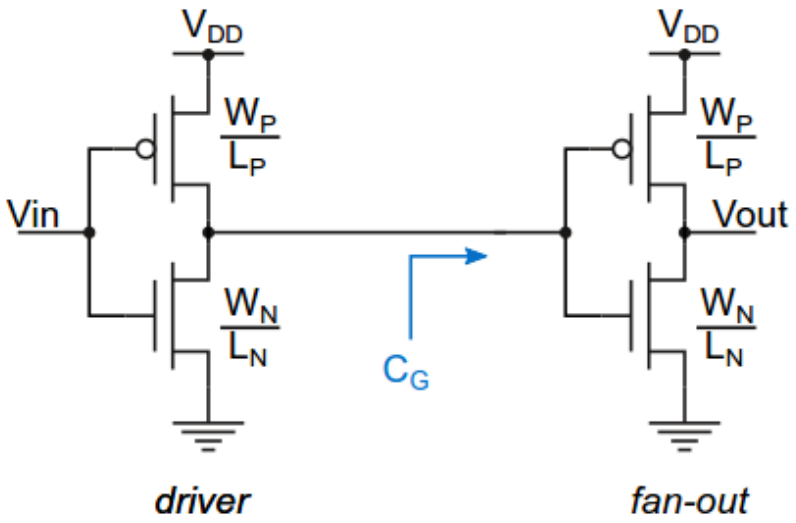
$$t_p = C\frac{Vdd/2}{\frac{1}{2}K\left(V_{DD}-V_T\right)^2}$$

Capacità di carico

Come sappiamo la capacità di carico è dovuta essenzialmente alla capacità parassita della linea di interconnessione (che collega l'invertitore al suo fan-out) ed alla capacità offerta dal fan-out stesso. La Figura mostra il semplice caso di un invertitore che pilota un fan-out unitario:



Capacità di carico



L'uscita del driver comanda sia il NMOS che il PMOS dello stadio successivo. La capacità offerta dal fan-out può essere stimata come somma delle capacità di gate dei dispositivi pilotati:

$$C_G=(W_N L_N C'ox) + (W_P L_P C'ox)$$

Scelta della *L* dei MOS

$$C_G=(W_N L_N C'ox) + (W_P L_P C'ox)$$

$$t_p = C \frac{Vdd/2}{\frac{1}{2} K \left(V_{DD} - V_T\right)^2}$$

Per ridurre il t_p è necessario aumentare il K del driver e ridurre la C_G del fan-out: **entrambi i risultati si ottengono minimizzando la lunghezza di canale dei dispositivi**. Pertanto **in logica CMOS tutti i dispositivi vengono realizzati con la minima lunghezza di canale che la tecnologia costruttiva è in grado di ottenere**.

Scelta della *W* dei MOS

$$t_p = C\frac{Vdd/2}{\frac{1}{2}K\left(V_{DD}-V_T\right)^2}$$

$C_G=(W_N L_N C'ox) + (W_P L_P C'ox)$

Mentre la scelta della L è immediata, lo stesso non può dirsi per la W dei dispositivi. Aumentando la W aumenta K (e ciò tende a ridurre i tempi di propagazione) ma al contempo aumenta la capacità carico (peggiora il ritardo e, come vedremo fra breve, la potenza dissipata).

In definitiva, per un invertitore CMOS simmetrico abbiamo un solo grado di libertà, legato alla scelta di W_N , infatti:

W_P = 2.5 W_N (per ottenere caratteristiche simmetriche)

L_N=L_P = dimensione minima offerta dalla tecnologia

Esempio 1

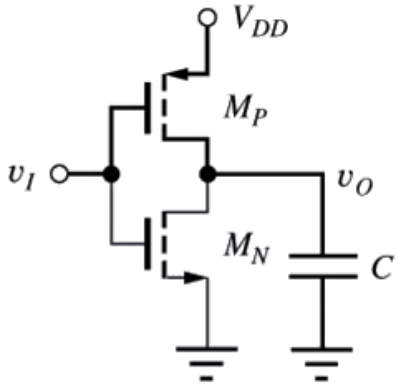
- Dimensionare un invertitore in modo da ottenere un ritardo di 250ps per un carico di 0.1pF. Assumere:

$$V_{DD} = 3.3V \qquad K_n^{'} = 100\frac{\mu A}{V^2}$$

$$C = 0.1pF$$

$$V_{TN} = -V_{TP} = 0.75V \qquad K_p^{'} = 40\frac{\mu A}{V^2}$$

Esempio 1



- Imponiamo dapprima la simmetria del circuito:

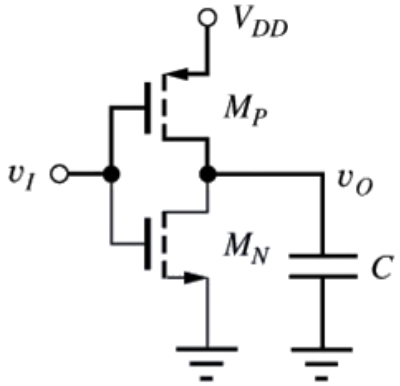
$$K_n'\left(\frac{W}{L}\right)_n = K_p'\left(\frac{W}{L}\right)_p$$

- La lunghezza di canale L è la minima consentita dalla tecnologia

$$L_n = L_p = L_{min}$$

- Pertanto: Wp=2.5 Wn

Esempio 1



- W_n (o W_p) è ottenuta imponendo il vincolo su t_p (notare che C è assegnata):

$$t_{phl} = C \cdot \frac{V_{DD}/2}{\frac{K_n'}{2} \cdot \left(\frac{W}{L}\right)_n \cdot \left(V_{DD} - V_{TN}\right)^2}$$

$$\left(\frac{W}{L}\right)_n = C \cdot \frac{V_{DD}/2}{\frac{K_n'}{2} \cdot t_{phl} \cdot \left(V_{DD} - V_{TN}\right)^2} = 2.03 \longrightarrow$$

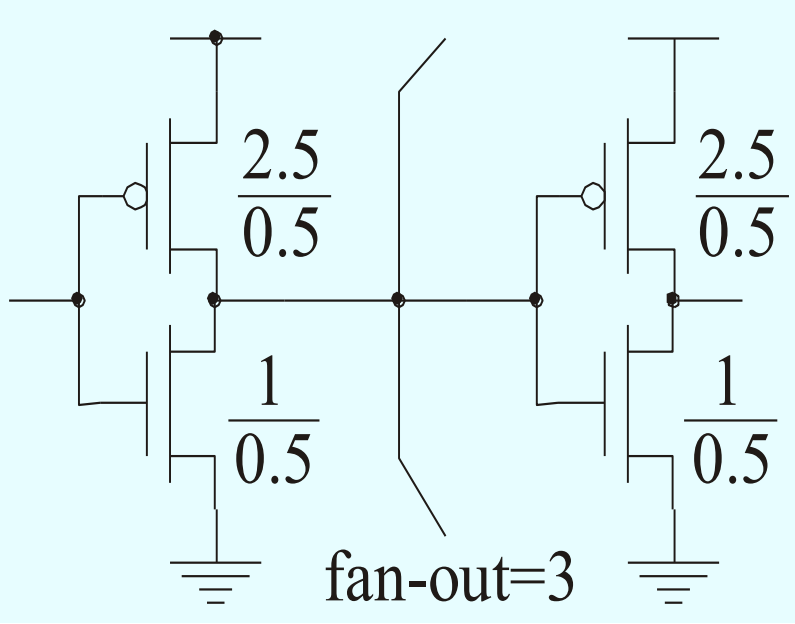
per L_{min}=100nm:

$W_n = 203\text{nm}$

$W_p = 507\text{nm}$

Esempio 2

Calcolare il ritardo dell'invertitore in Figura, che ha come carico 3 invertitori identici.



V_{DD}=3.3V
K'_N=75 μA/V^2
$K'_P = K'_N/2.5$
V_{TN} = 0.8V
$V_{TP}=|V_{Tn}|$,
C'_{OX}=4 fF/μm^2
L=0.5μm

Esempio 2

La capacità di carico, tenendo conto che il fan-out è 3, è data da:

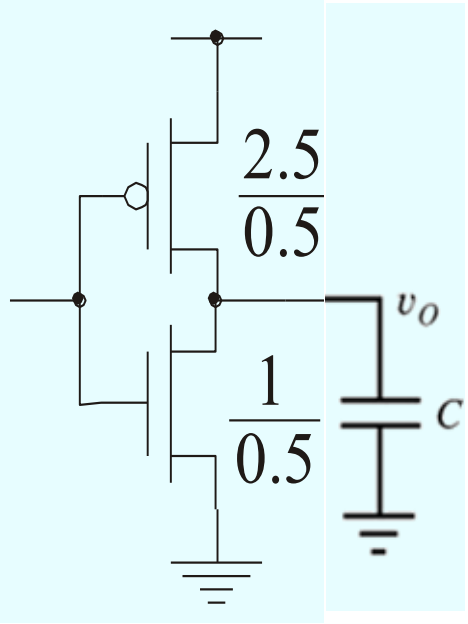
$$Ctot = 3\left(WnLn + WpLp\right)C'ox = 21\,fF$$

Il circuito è simmetrico, poiché: $V_{TN}=|V_{TP}|=V_T$; Kn=Kp; (il rapporto W/L del PMOS è 2.5 volte W/L del NMOS e compensa pertanto le differenze di mobilità)

$$t_p = \frac{CtotV_{DD}/2}{\frac{K}{2}\left(V_{DD}-V_T\right)^2} = 37\,ps$$

Esempio 3

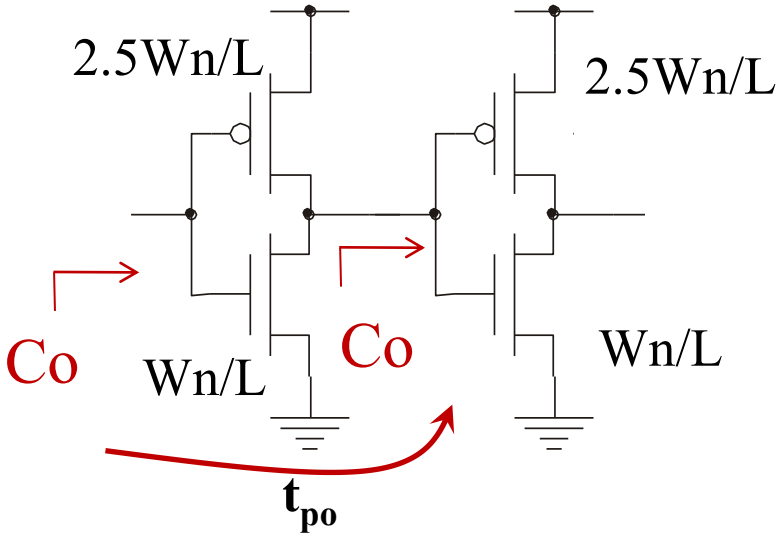
Consideriamo l'invertitore dell'esempio 1 e supponiamo che debba pilotare una capacità di carico di 2.1pF. Determinare il tempo di propagazione.



Poiché la capacità è 100 volte maggiore rispetto all'esempio 2, il tempo di propagazione sarà anch'esso 100 volte maggiore:
t_P=3.7 ns

Stadi separatori (buffer)

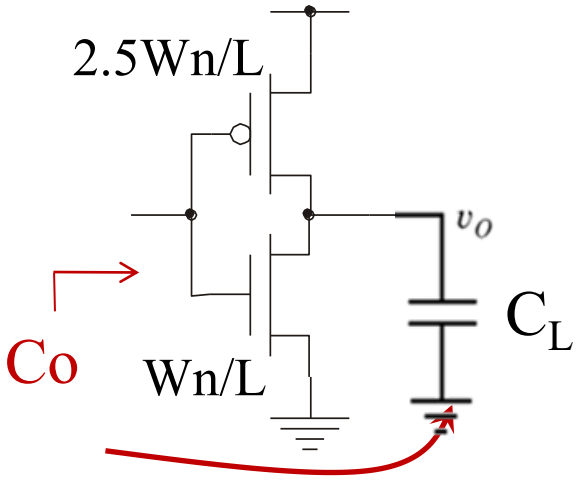
Il ritardo di un invertitore (come di qualsiasi porta CMOS) è direttamente proporzionale alla capacità di carico.
Il ritardo minimo di un invertitore è quello manifestato quando la capacità di carico è minima, ovvero quando il fanout è unitario.
Consideriamo dunque due invertitori identici in cascata:



Indicando con Co la capacità di ingresso del primo invertitore, la capacità di carico sarà anch'essa Co.
Chiamiamo $\mathbf{t_{po}}$ il ritardo dell'invertitore con fanout unitario

Stadi separatori (buffer)

Supponiamo ora che l'invertitore debba pilotare una capacità di carico C_L>>Co. Questa condizione si realizza, ad esempio, negli stadi di uscita di un circuito integrato, che devono pilotare dei carichi esterni al chip, caratterizzati da una capacità molto maggiore di quelle che troviamo in un circuito integrato



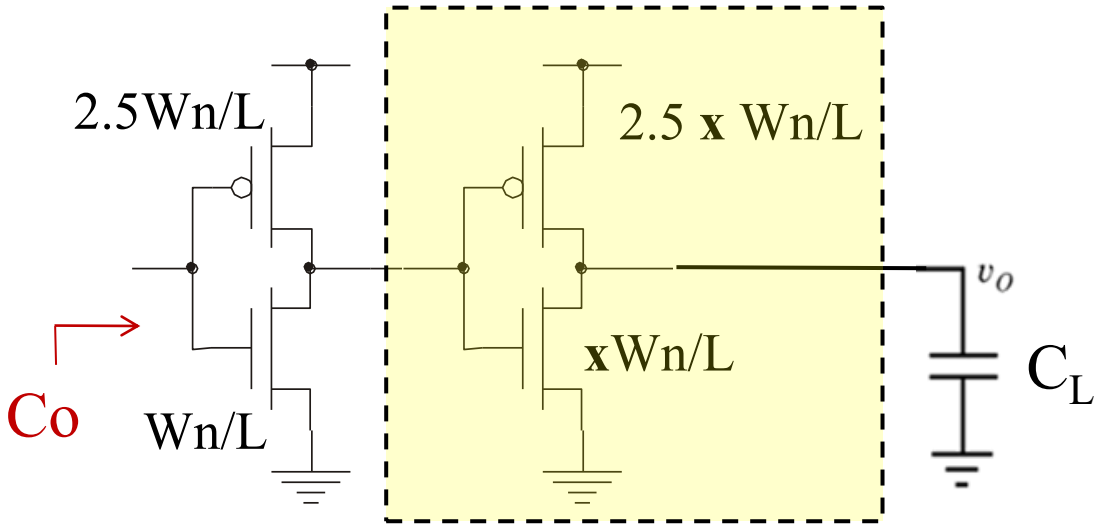
$\mathbf{t_p >> t_{po}}$

Il tempo di propagazione (essendo proporzionale alla capacità di carico) sarà:

$\mathbf{t_P = t_{po}\,(C_L/Co) >> t_{po}}$

Stadi separatori (buffer)

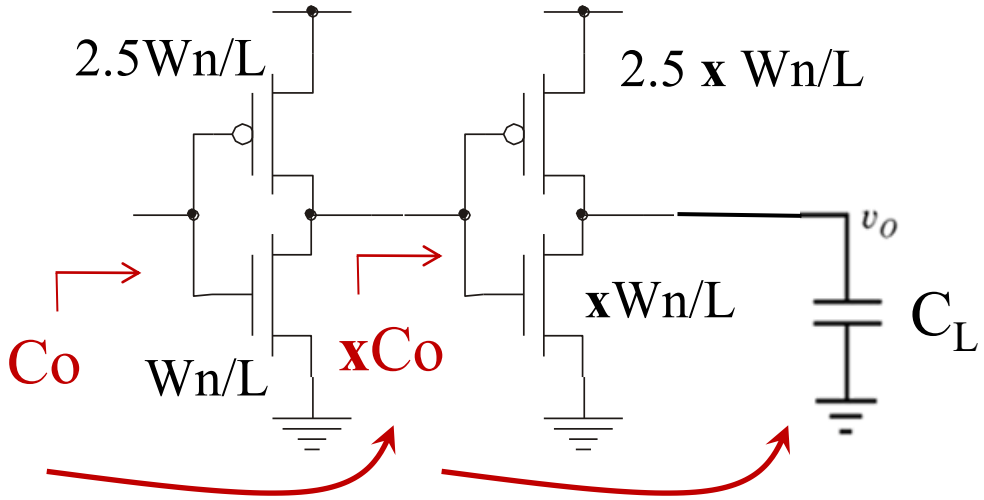
Al fine di ridurre il tempo di propagazione complessivo, interponiamo fra l'invertitore ed il carico un altro inverter, di dimensioni **x** volte maggiori. Questo invertitore aggiunto rappresenta un esempio di stadio separatore o buffer.



buffer

Il ritardo complessivo può essere minimizzato, agendo sulle dimensioni del buffer, ovvero sul parametro **x**

Stadi separatori (buffer)



x t_{po} **t_{po} C_L/(xCo)**

Il ritardo del primo inverter è **xt_{po}** (la capacità di carico è **x** volte maggiore di Co**)**

Il ritardo del secondo invertitore è: **t_{po} (C_L/xCo)** (l'invertitore è più grande di un fattore x rispetto al primo, ma la capacità di carico è più grande di un fattore **C_L/Co**

$t_{P,tot}$= xt_{po} + t_{po} (C_L/xCo) = f(x)

Derivando rispetto ad x e ponendo la derivata a zero si ottiene la condizione che minimizza il ritardo.

Stadi separatori (buffer)

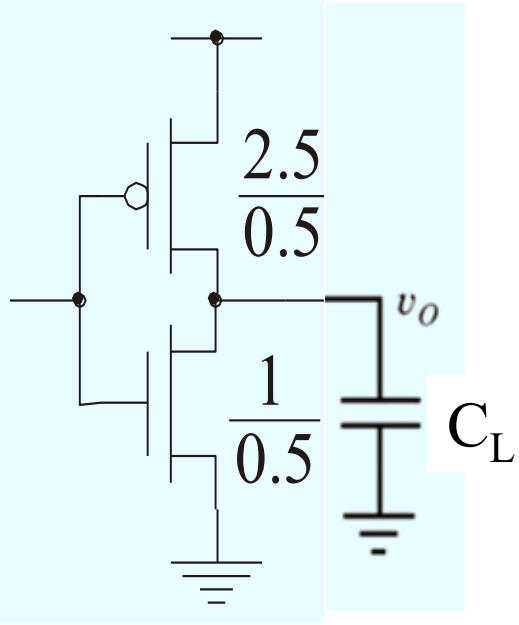
$\mathbf{d\ f/dx = t_{po} - t_{po}\ (C_L/x^2Co) = 0}$ $\Rightarrow$ $xopt = \sqrt{C_L / Co}$

Con il valore ottimale di x, si ha: $t_{P,opt} = 2tpo\sqrt{C_L / Co}$

Da notare che, all'ottimo, i tempi di propagazione dei due invertitori sono uguali fra loro e pari a: $tpo\sqrt{C_L / Co}$

Stadi separatori (Esempio)

Consideriamo l'esempio 3 precedente, con C_L=2.1pF



$$Co = \left(WnLn + WpLp\right)C'ox = 7\,fF$$

$$t_{p0} = \frac{CoV_{DD}/2}{K\left(V_{DD} - V_T\right)^2} = 12.3\,ps$$

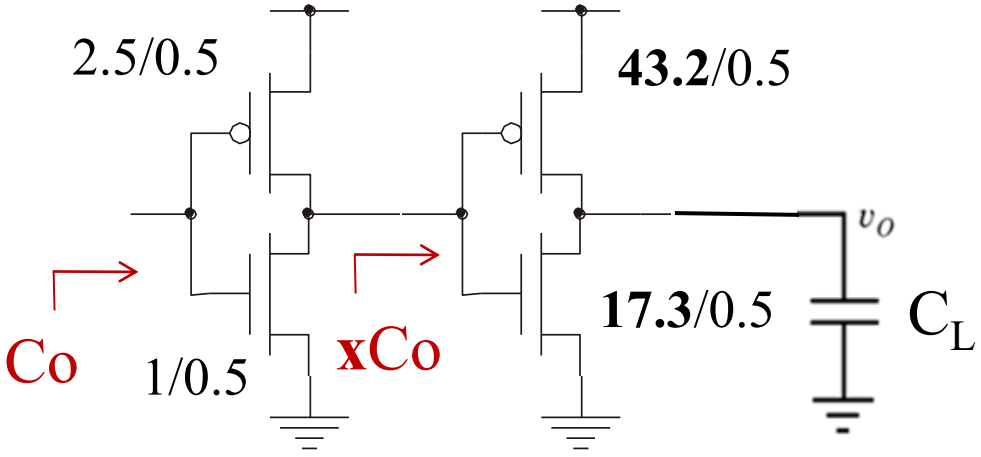
In assenza di buffer:
tp=tpo CL/Co= 3.7 ns

Stadi separatori (Esempio)

Introducendo il buffer: $xopt = \sqrt{C_L / Co} = 17.3$

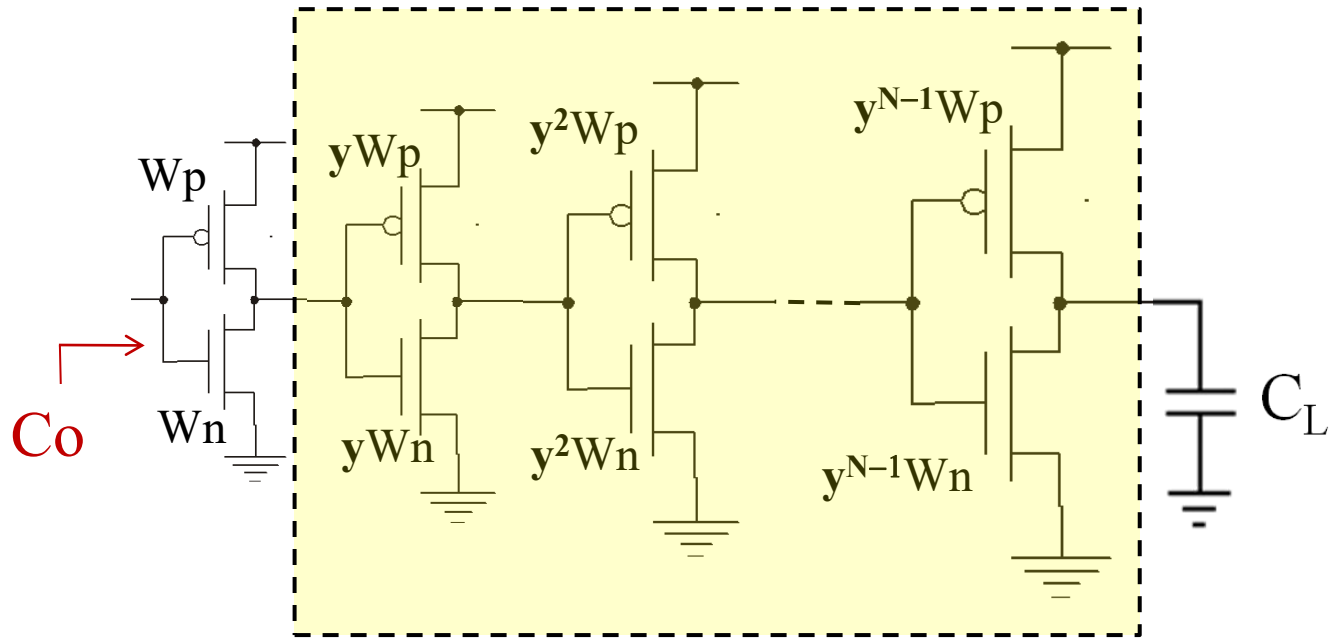
$$t_{P,opt} = 2tpo\sqrt{C_L / Co} = 430\,ps$$

il ritardo complessivo si è ridotto di più di 8 volte



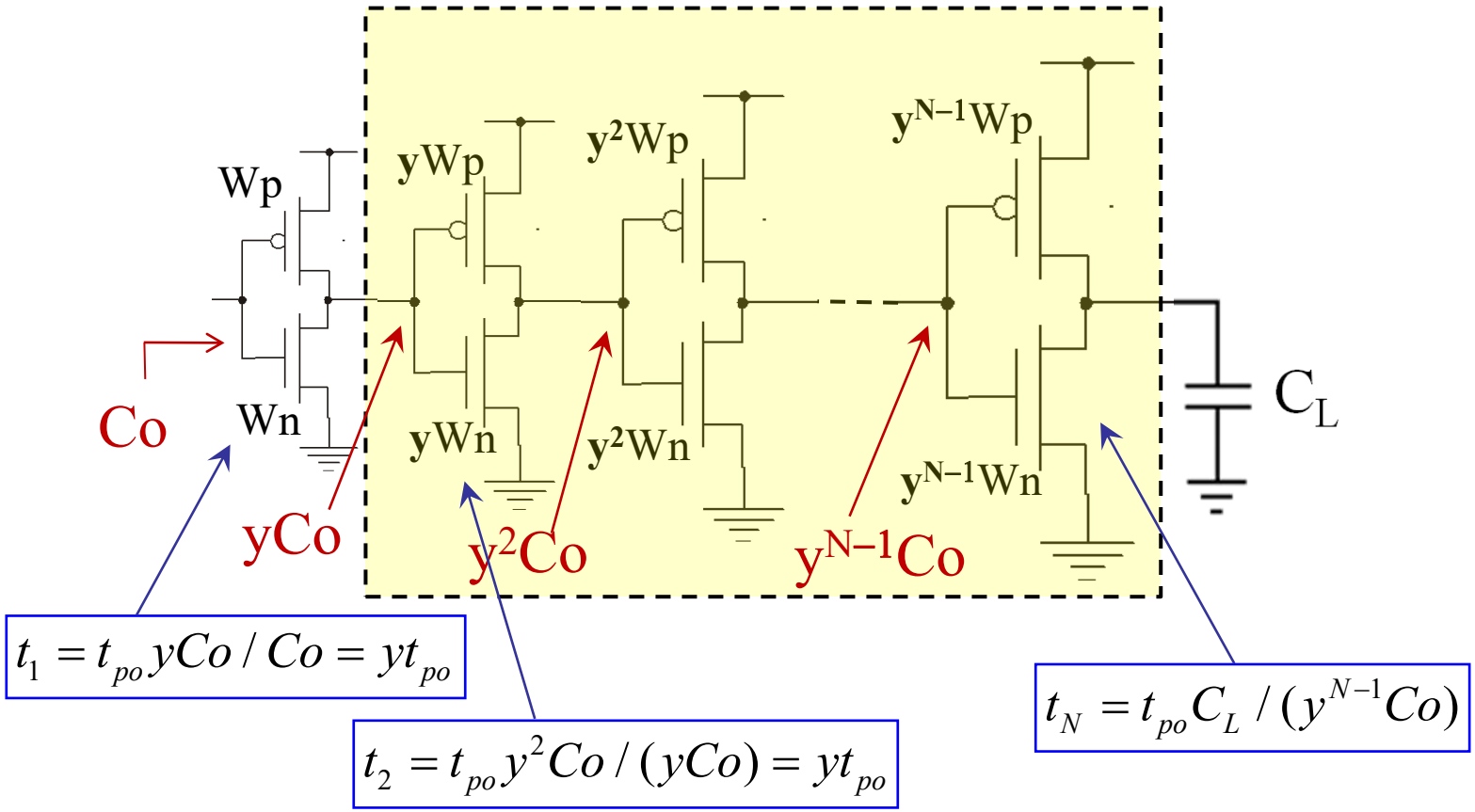
Buffer ottimo

Una versione ulteriormente ottimizzata del buffer è costituita inserendo **N**-1 invertitori (oltre a quello iniziale), ognuno più grande del precedente di un fattore **y**



Buffer ottimo

Valutiamo le capacità di ingresso dei vari invertitori ed i ritardi



Buffer ottimo

$$t_1 = yt_{po}$$

$$t_2 = yt_{po}$$

$$\ldots$$

$$t_{N-1} = yt_{po}$$

$$t_N = t_{po} C_L / (y^{N-1} Co)$$

imponendo:

$$C_L / (y^{N-1} Co) = y$$

rendiamo uguali i ritardi di tutti gli stadi.

Deve essere:

$$y^N = C_L / Co \quad \Rightarrow \quad N \ln(y) = \ln(C_L / Co) \quad \Rightarrow \quad N = \frac{\ln(C_L / Co)}{\ln(y)}$$

Buffer ottimo

Il ritardo complessivo è: $t_{TOT} = t_1 + t_2 + \ldots + t_N = Nyt_{po}$

Sostituendo $N = \dfrac{\ln(C_L / Co)}{\ln(y)}$

$$t_{TOT} = \frac{y}{\ln(y)} t_{po} \ln(C_L / Co)$$

Derivando rispetto ad y ed azzerando la derivata otteniamo il valore ottimo di y:

$$\frac{dt_{TOT}}{dy} = \frac{\ln(y) - y/y}{\ln^2(y)} t_{po} \ln(C_L / Co) = 0 \quad \Rightarrow \quad \ln(y) = 1 \quad \Rightarrow y = e \simeq 2.718\ldots$$

Buffer ottimo

Dunque, il fattore di incremento ottimo di ogni invertitore è:

$$y_{opt} = e \simeq 2.718...$$

Il numero ottimo di stadi è:

$$Nopt = \ln(C_L / Co)$$

Il ritardo ottimo è:

$$t_{TOT,OPT} = et_{po} \ln(C_L / Co)$$

Stadi separatori (Esempio)

Consideriamo l'esempio precedente, con C_L=2.1pF, Co=7fF

$$Nopt = \ln(C_L / Co) = 5.7 \rightarrow 6$$

ovviamente il numero di invertitori deve essere intero! Si devono aggiungere 5 invertitori allo stadio iniziale

$$t_{TOT,OPT} \approx et_{po} Nopt \simeq 200\,ps$$

In assenza di buffer: tp=tpo CL/Co= 3.7 ns

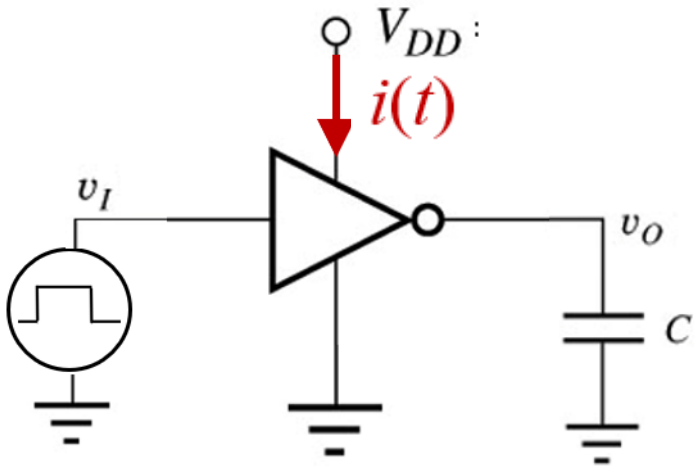
Aggiungendo un unico invertitore: $t_{P,opt} = 2tpo\sqrt{C_L / Co} = 430\,ps$

Dissipazione di Potenza

Dissipazione di Potenza

- La dissipazione di potenza statica nelle logiche CMOS è molto piccola ed è dovuta alle correnti di perdita che fluiscono attraverso I MOS che sono nominalmente interdetti.
- Le correnti di perdita (*leakage*) sono dovute alle giunzioni drain-substrato (diodi inversamente polarizzati che hanno una corrente di saturazione inversa non nulla) ed alla *corrente sotto-soglia* che fluisce fra drain e source anche quando $V_{GS}=0$.

Dissipazione di potenza dinamica



Consideriamo un invertitore CMOS, sottoposto ad un ingresso che produce delle commutazioni nel circuito, con un periodo T.

Il generatore di tensione V_{DD}, che alimenta il circuito, eroga una corrente $i(t)$

La *potenza istantanea* erogata dal generatore e dissipata nell'invertitore è data da: $p(t) = V_{DD}\, i(t)$

Il parametro più significativo è **la potenza media** erogata dal generatore e dissipata nell'invertitore, data da:

$$P_D = \frac{1}{T} V_{DD} \int_T i(t)dt$$

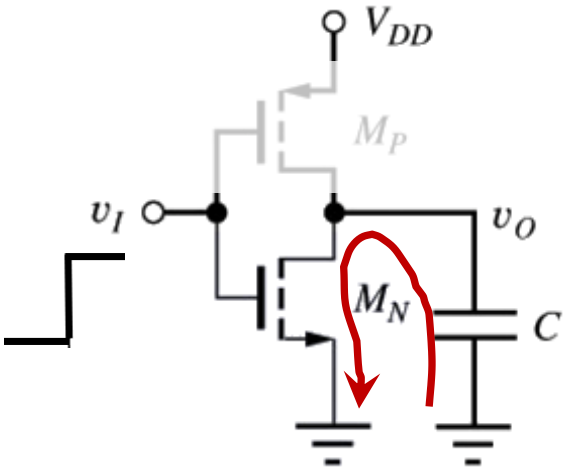
Dissipazione di potenza dinamica

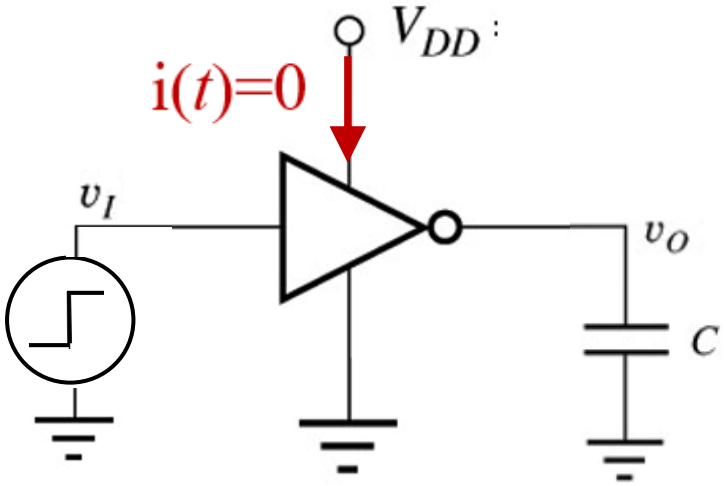
- La dissipazione di potenza dinamica è dovuta essenzialmente all'energia necessaria per la carica/scarica della capacità di uscita.

- Vi è anche un ulteriore contributo (generalmente trascurabile, che quindi non consideriamo nel seguito) legato alla fase di conduzione simultanea di NMOS e PMOS nel breve intervallo in cui il segnale d'ingresso è compreso fra V_T e V_{DD}-V_T.

Dissipazione di potenza dinamica

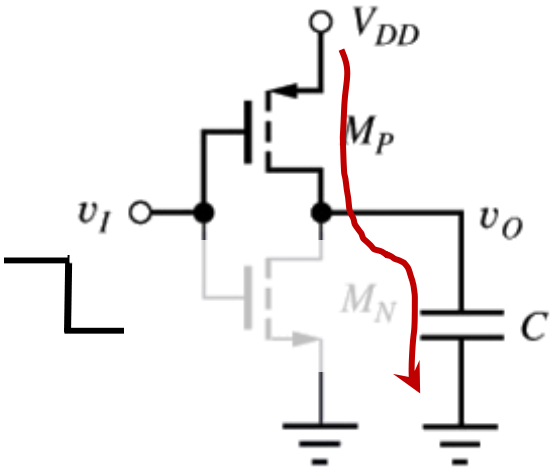
Durante la transizione 1→0 dell'uscita, la capacità di uscita C si scarica attraverso il NMOS. In questa fase **non c'è erogazione di corrente** da parte dell'alimentatore: i(t)=0

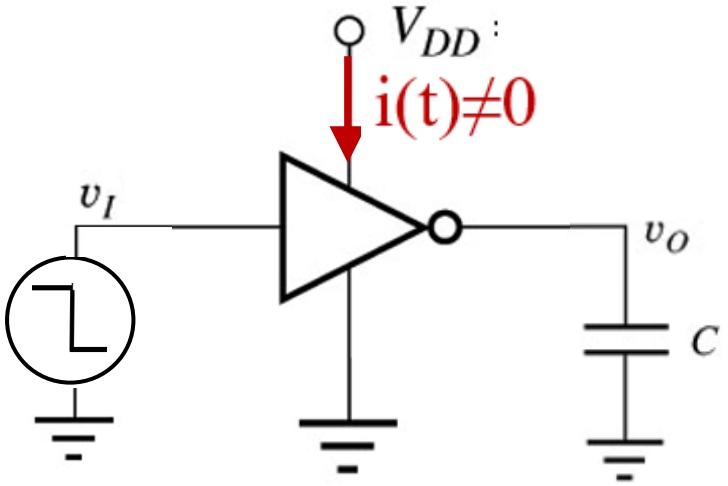




Dissipazione di potenza dinamica

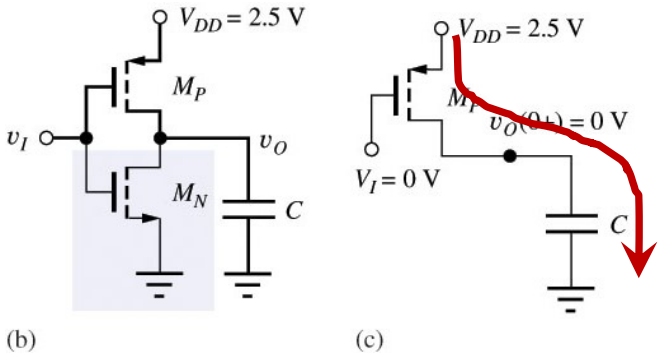
Durante la transizione 0→1 dell'uscita la capacità di uscita C si carica attraverso il PMOS. In questa fase c'è una erogazione di corrente (e quindi una dissipazione di potenza) da parte dell'alimentatore: i(t)≠0

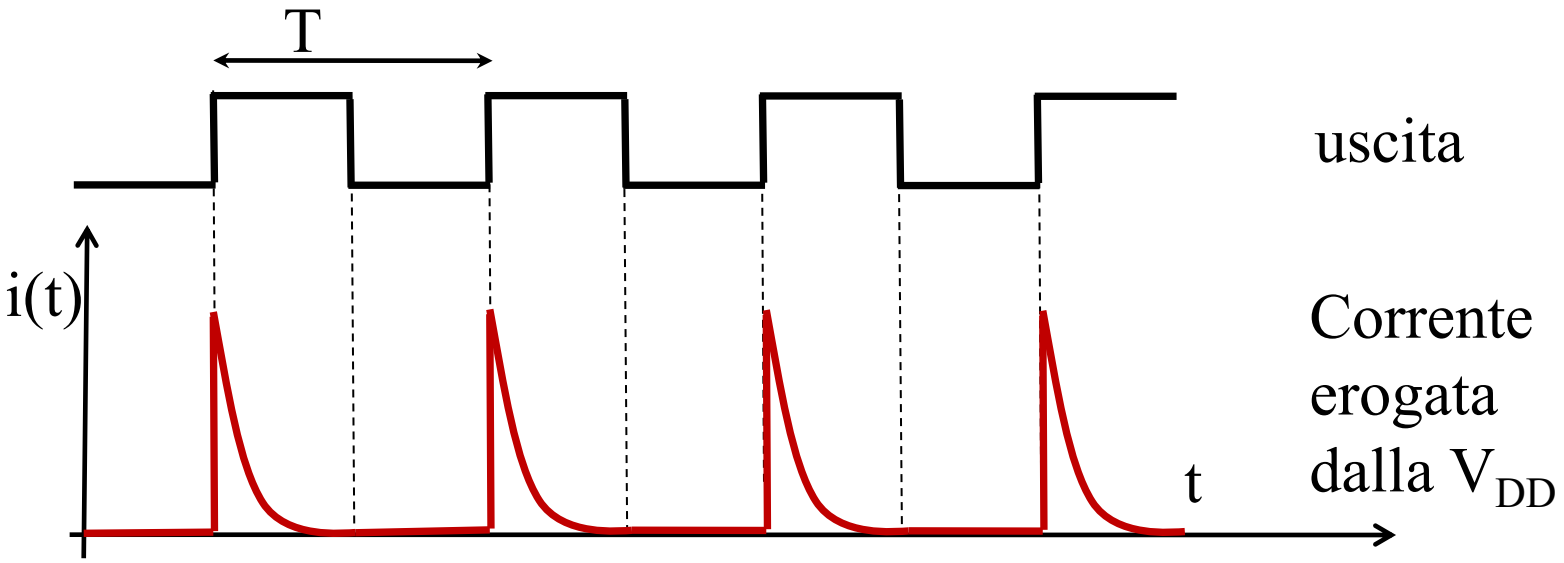




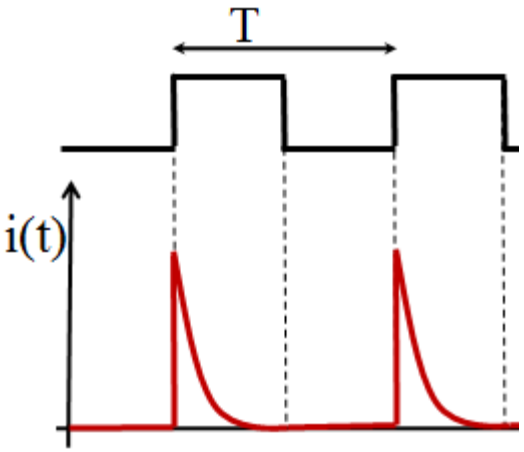
Dissipazione di potenza dinamica

Supponiamo di avere un segnale periodico, di periodo T, in ingresso all'invertitore





Dissipazione di potenza dinamica



La potenza media erogata dall'alimentatore in un periodo è data da:

$$P_D = \frac{1}{T} V_{DD} \int_T i(t)dt$$

L'integrale si può estendere al primo semiperiodo del segnale (la fase di carica della C).

La corrente è quella che scorre nella capacità di uscita: $i = C\frac{dv_O}{dt}$

$$P_D = \frac{1}{T} V_{DD} \int_0^{T/2} C\frac{dv_O}{dt}dt = \frac{1}{T} CV_{DD} \int_0^{V_{DD}} dv_O$$

Dissipazione di potenza dinamica

$$P_D = \frac{1}{T} C V_{DD} \int_0^{V_{DD}} dv_O = \frac{1}{T} C V_{DD}^2$$

$$P_D = f \cdot C \cdot V_{DD}^2$$

Dove *f* è la frequenza del segnale d'ingresso.

Si noti che la potenza P_D viene **erogata** dall'alimentatore e viene **dissipata** sul PMOS e sul NMOS

Dynamic Voltage and Frequency scaling

Ricordiamo che i ritardi di propagazione di un circuito CMOS dipendono dalla tensione di alimentazione ed in particolare si riducono al decrescere della tensione di alimentazione:

$$t_p = \frac{CV_{DD}/2}{\frac{K}{2}\left(V_{DD} - V_T\right)^2}$$

Abbiamo dunque due esigenze contrastanti: al fine di massimizzare la velocità dei circuiti è opportuno aumentare la tensione di alimentazione, mentre è opportuno diminuirla per contenere la potenza dissipata.

Nei microprocessori attuali si cerca di **adattare la tensione di alimentazione alla frequenza di funzionamento richiesta**, in modo da ottimizzare la dissipazione di potenza

Dynamic Voltage and Frequency scaling

Quando è richiesto un uso intensivo della CPU si aumenta la tensione di alimentazione, che viene invece ridotta nei momenti in cui non sono richieste le massime prestazioni al circuito.

Questa tecnica viene denominata *Dynamic Voltage and Frequency scaling*

Esempio

Un microprocessore, alimentato alla tensione di Vdd=1.25V, opera ad una frequenza massima di f=2GHz con una potenza dissipata Pd=30W.
A quale tensione è necessario alimentare il circuito per portare la frequenza massima a 3GHz?
Quale sarà la nuova dissipazione di potenza?
Ipotizzare V_T=0.5V.

Esempio (cont.)

La frequenza massima di funzionamento è inversamente proporzionale al ritardo delle porte logiche che lo compongono. Si ha dunque:

$$f \propto \frac{1}{t_p}; \quad \text{con:} \quad t_p = \frac{CV_{DD}/2}{\frac{K}{2}\left(V_{DD} - V_T\right)^2}$$

Risulta, quindi: $f \propto \dfrac{\left(V_{DD} - V_T\right)^2}{V_{DD}}$

Questa relazione, ottenuta per un semplice invertitore, vale anche per un sistema molto più complesso come un microprocessore

Esempio (cont.)

Si può pertanto scrivere: $$f = A\frac{\left(V_{DD} - V_T\right)^2}{V_{DD}}$$

La costante di proporzionalità, A, si può ricavare dai dati del problema:

$$A = \frac{fV_{DD}}{\left(V_{DD} - V_T\right)^2} = 4.44GHz/V$$

Risulta, sostituendo i valori numerici:

$$f = 4.44\frac{\left(V_{DD} - 0.5\right)^2}{V_{DD}}$$

Esempio (cont.)

Imponiamo ora il nuovo valore di f e valutiamo la corrispondente tensione di alimentazione:

$$3 = 4.44\frac{\left(V'_{DD} - 0.5\right)^2}{V'_{DD}} \quad \Rightarrow \quad \left(V'_{DD} - 0.5\right)^2 = \frac{3}{4.44}V'_{DD}$$

Si ha una equazione di secondo grado in V'_{DD} che può essere facilmente risolta: $V'_{DD}=0.838\pm0.672$V. Delle due soluzioni l'unica con senso fisico è quella che fornisce un valore di tensione superiore a quella di partenza. Si ha quindi:

$V'_{DD}=1.51$V

Esempio (cont.)

Per il calcolo della potenza dissipata, notiamo che: $P = fC\left(V_{DD}\right)^2$

Aumentando tensione di alimentazione e frequenza risulta:

$$P' = f'C\left(V_{DD}'\right)^2$$

Si ha quindi: $$\frac{P'}{P} = \frac{f'}{f}\left(\frac{V_{DD}'}{V_{DD}}\right)^2$$

Sostituendo i valori numerici:

P'=2.188P=65.7W

Multicore

Negli ultimi anni, la frequenza massima dei microprocessori si è assestata intorno a qualche GHz, nonostante il continuo progredire della tecnologia CMOS.
Questo fenomeno è legato alla necessità di limitare la dissipazione di potenza dei microprocessori.
Al fine di aumentare comunque le capacità elaborative (senza aumentare la frequenza di clock) sono divenuti di uso comune sistemi multi processore (multi-core).
Tali sistemi, infatti, a parità di potenza elaborativa, **sono potenzialmente più efficienti da un punto di vista energetico** rispetto ad un sistema con una singole CPU.

Esempio 2

Un microprocessore opera ad un frequenza di 4GHz dissipando una potenza di 80W. La tensione di alimentazione è: V_{DD}=1.2V, la tensione di soglia è: V_T=0.4V.

Consideriamo una struttura dual-core, con le stesse capacità elaborative. Ogni core operera' ad una frequenza di 2GHz (invece di 4GHz).

Valutare la tensione di alimentazione a cui si può far operare il processore dual-core e la nuova dissipazione di potenza.

Ripetere nel caso di sistema quad-core, con ogni core operante ad 1GHz.

Esempio 2 (continua)

Nella struttura dual-core, supponiamo inizialmente di alimentare entrambi i cores alla tensione di alimentazione iniziale, V_{DD}=1.2V

Visto che ogni core opera a frequenza dimezzata rispetto al caso iniziale, la potenza dissipata da ogni core sarà:

$P_{CORE}=CV^2_{DD}(f/2) = 40W$

La potenza complessiva è: $P_{TOT}=2P_{CORE}= 80W$

Operando alla stessa tensione di alimentazione, non c'è alcun vantaggio in termini di riduzione di potenza!

Esempio 2 (continua)

Visto che ogni core dovrà operare ad una frequenza inferiore rispetto ai 4GHz iniziali, la nostra strategia sarà quella di **ridurre la tensione di alimentazione**. In questo modo avremo un sensibile risparmio di potenza, a parità di capacità elaborative.

Esempio 2 (continua)

Sappiamo che: $f \propto \frac{1}{t_p}$; con: $t_p = \frac{CV_{DD}/2}{\frac{K}{2}(V_{DD}-V_T)^2}$

possiamo quindi scrivere: $f \propto \frac{(V_{DD}-V_T)^2}{V_{DD}}$

Ovvero: $f = A\frac{(V_{DD}-V_T)^2}{V_{DD}}$

La costante di proporzionalità, A, si può ricavare dai dati del problema:

$$A = \frac{fV_{DD}}{(V_{DD}-V_T)^2} = 7.5GHz/V$$

Esempio 2 (continua)

Si ha dunque: $$f = 7.5\frac{(V_{DD} - 0.4)^2}{V_{DD}}$$

Risolvendo l'equazione precedente in V_{DD}, possiamo calcolare la tensione di alimentazione che ci consente di operare ad f=2GHz.

Si ha: **$V'_{DD} = 0.89V$**

La potenza dissipata da ogni core, essendo proporzionale a: $f\,V^2_{DD}$ diviene:

Pcore / Piniziale = $(1/2)\,(0.89/1.2)^2 = 0.275$

Si ha quindi: Pcore= $80 \times 0.275 = 22W$

Avendo due core: **Ptot(dual-core)=44W**

La dissipazione di potenza si riduce circa del 50%

Esempio 2 (continua)

Nel caso quad-core, ogni core può operare ad 1GHz; Risolvendo l'equazione precedente in V_{DD}, possiamo calcolare la nuova tensione di alimentazione:

Si ha: $\mathbf{V''_{DD} = 0.71V}$

La potenza dissipata da ogni core, essendo proporzionale a: $f\,V^2_{DD}$ diviene:

Pcore / Piniziale = $(1/4)\,(0.71/1.2)^2 = 0.0875$

Si ha quindi: Pcore= $80 \times 0.0875 = 7W$

Avendo quattro core: **Ptot(quad-core)=28W**

Riduzione in scala dei circuiti CMOS

Scaling CMOS

- Negli anni si è assistito ad un continuo miglioramento della risoluzione dei processi tecnologici utilizzati per la realizzazione dei circuiti CMOS.
- La riduzione delle dimensioni dei dispositivi produce un miglioramento delle prestazioni e la possibilità di integrare funzioni sempre più complesse su di un unico chip.
- Passando da una tecnologia ad un'altra più avanzata si effettua una **riduzione in scala** (**scaling**) dei dispositivi. Ad esempio, passando da una tecnologia con lunghezza minima di canale L=90nm ad una più avanzata con L=60nm possiamo ridurre tutte le dimensioni geometriche di un fattore **x**=9/6
- Quali sono le conseguenze dello scaling?

Scaling a campo costante

- Assumiamo che con il progredire della tecnologia, tutte le dimensioni geometriche si riducono di un fattore **x>1** e contemporaneamente, anche la tensione di alimentazione del circuito e la tensione di soglia dei dispositivi si riduce dello stesso fattore.
- In questo modo **i campi elettrici presenti nei dispositivi restano invariati (scaling a campo costante).**
- Se si riducessero solo le dimensioni geometriche, i campi elettrici nel dispositivo aumenterebbero con la conseguente possibilità di distruggere il dispositivo.

Scaling a campo costante

- Le grandezze geometriche si riducono di **x**

  $L \rightarrow L/\mathbf{x}$

  $W \rightarrow W/\mathbf{x}$

  $tox \rightarrow tox/\mathbf{x}$

- Le tensioni si riducono di **x**:

  $V_{DD} \rightarrow V_{DD}/\mathbf{x}$

  $V_T \rightarrow V_T/\mathbf{x}$

Scaling a campo costante

- Vediamo cosa accade per gli altri parametri:

  $C'ox=\varepsilon ox/tox \rightarrow \mathbf{x}\ C'ox$ (la capacità di gate per unità di area)

  $Cg=C'ox\ W\ L \rightarrow C'ox / \mathbf{x}$ (capacità di gate)

  $K'=\mu C'ox \rightarrow \mathbf{x}\ K'$

  $K=K'W/L \rightarrow \mathbf{x}\ K$

  $Imax=0.5K(V_{DD}-V_T)^2 \rightarrow Imax / \mathbf{x}$ (corrente in pinch-off)

Valutiamo ora il tempo di propagazione:

$$\mathbf{tp \approx (V_{DD}/2)Cg/Imax \rightarrow tp / x}$$

Il ritardo si riduce (migliora) del fattore di scaling x

La frequenza a cui può funzionare il circuito aumenta di x

$$\mathbf{f \rightarrow f\ x}$$

Scaling a campo costante

Vediamo la potenza dissipata da una singola gate:

$$P_D \approx f\, V^2_{DD} Cg \rightarrow P_D / x^2$$

La potenza dissipata da un intero circuito si può stimare osservando che, *a parità di area complessiva, il numero di gate aumenta di un fattore* x^2. Dunque, la potenza complessiva tende a rimanere costante:

$$P_{D,tot} \rightarrow P_{D,tot}$$

Margini di rumore: riducendo tutte le tensioni si riducono del fattore **x**:

$$N_M \rightarrow N_M / x$$

Scaling a campo costante

Riassumendo:

$$tp \rightarrow tp / x$$

$$f \rightarrow f\, x$$

$$P_D \rightarrow P_D / x^2$$

$$P_{D,tot} \rightarrow P_{D,tot}$$

$$N_M \rightarrow N_M / x$$

Scaling a campo costante

In pratica, non si riesce ad effettuare uno scaling a campo costante, in quanto **non è possibile ridurre la tensione dello fattore di cui si riducono le tensioni geometriche** (alcune grandezze, come la V_T, non possono essere modificate a piacimento). **La tensione scala meno fortemente del parametro geometrico, con conseguente aumento della potenza dissipata totale**.

Da queste considerazioni l'esigenza di sviluppare le tecniche di *Dynamic Voltage and Frequency scaling* ed i sistemi *multicore*, cui si è fatto cenno in precedenza**.**

Gli esempi seguenti mostrano cosa accade quando non è possibile effettuare lo scaling a campo costante.

Esempio 1 – scaling a campo costante

Un microprocessore, realizzato in una tecnologia a 90nm con V_{DD}=1.2V, V_T=0.4V, opera ad un frequenza di 1GHz dissipando una potenza di 10W.

Il sistema viene "portato" ad una tecnologia a 45nm. L'area complessiva del circuito (a causa di un piu' complesso set di istruzioni, all'introduzione di acceleratori grafici ecc) resta invariata rispetto alla versione precedente (il numero complessivo di gates integrate si quadruplica). Nella nuova tecnologia, la tensione di alimentazione e la tensione di soglia dei MOS si dimezza.

Valutare la nuova frequenza operativa e la nuova dissipazione di potenza del microprocessore.

Esempio 1 (continua)

Il fattore di scaling è **x**=2 e siamo nella condizione ideale di scaling a campo costante.

$$\mathbf{f \rightarrow f\ x = 2GHz}$$

$$\mathbf{P_{D,tot} \rightarrow P_{D,tot} = 10W}$$

La frequenza a cui può funzionare il microprocessore aumenta di un fattore due, mentre la potenza dissipata complessiva resta costante.

Esempio 2 – scaling a tensione costante

Un microprocessore, realizzato in una tecnologia a 90nm con V_{DD}=1.2V, V_T=0.4V, opera ad un frequenza di 1GHz dissipando una potenza di 10W.

Il sistema viene "portato" ad una tecnologia a 45nm. L'area complessiva del circuito (a causa di un piu' complesso set di istruzioni, all'introduzione di acceleratori grafici ecc) resta invariata rispetto alla versione precedente (il numero complessivo di gates integrate si quadruplica). **La tensione di alimentazione e la tensione di soglia dei MOS restano invariate nonostante lo scaling**.

Valutare la nuova frequenza operativa e la nuova dissipazione di potenza del microprocessore.

Esempio 2 (continua)

Il fattore di scaling è **x**=2. Non siamo in un caso di scaling a campo costante, poiché tensione di alimentazione e V_T restano invariate.

Si ha:

Imax=0.5K$(V_{DD}-V_T)^2$ → Imax **x** (a causa dell'aumento del K)

Tempo di propagazione:

$$\mathbf{tp \approx (V_{DD}/2)Cg/Imax \rightarrow tp \,/\, x^2}$$

(a causa dell'aumento di Imax e della riduzione di Cg)

La frequenza a cui può funzionare il circuito aumenta di x^2 e quindi: $f \rightarrow f\,x^2 = 4GHz$

Esempio 2 (continua)

La potenza di ogni singola gate del microprocessore diviene:

$\mathbf{P_D \approx f V^2_{DD} Cg \rightarrow P_D\, x}$ (f aumenta di x^2 mentre Cg si riduce di x)

La potenza totale diviene:

$\mathbf{P_{D,tot} \rightarrow x^3\, P_{D,tot} = 80W}$

Il nuovo processore è molto più veloce, ma la dissipazione di potenza cresce in maniera drammatica!