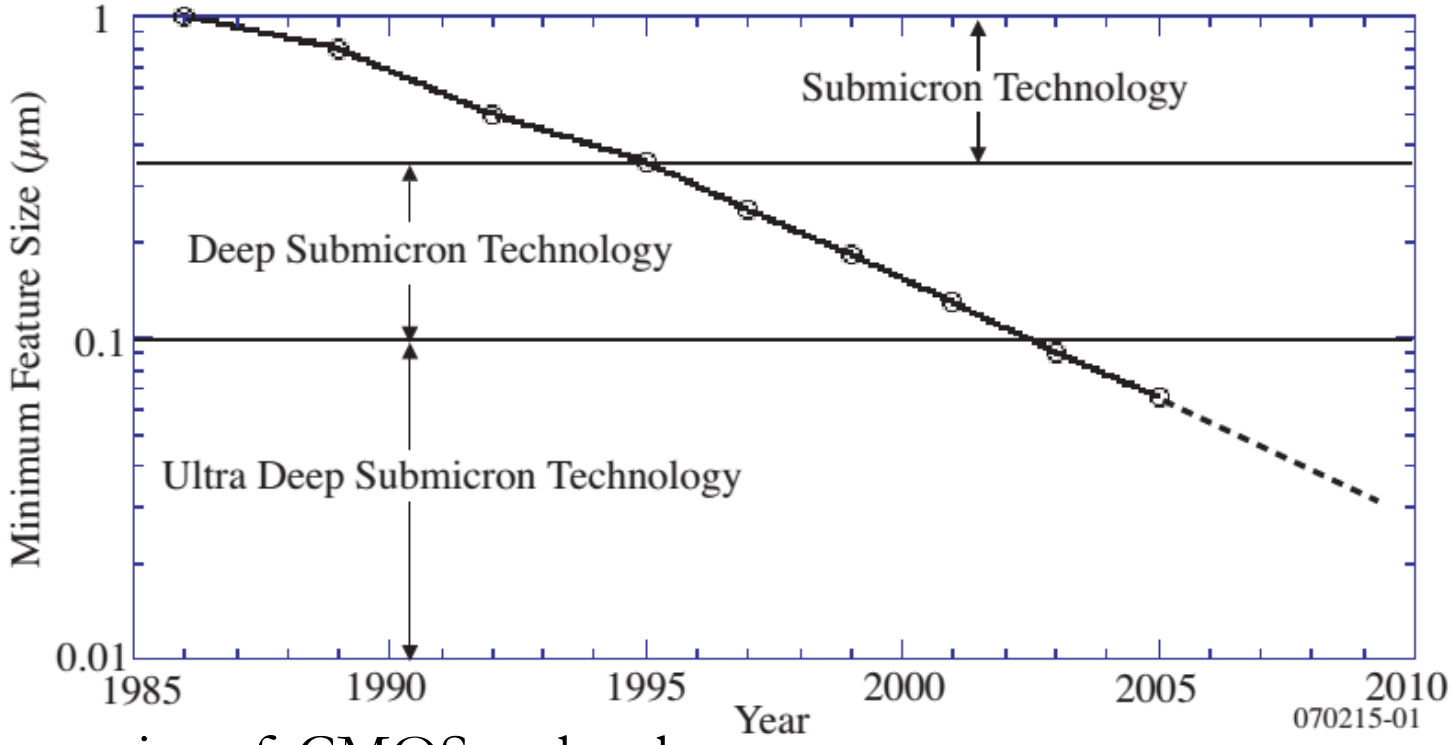
Scaling dei dispositivi

Deep and UltraDeep sub-micrometer MOS

Categorization of CMOS Technology

- **Minimum feature size as a function of time:**



Categories of CMOS technology:

- Submicron technology – Lmin > 0.35 microns
- Deep Submicron technology (DSM) – 0.1 μm ≤ Lmin ≤ 0.35 μm
- Ultra-Deep Submicron technology (UDSM) – Lmin ≤ 0.1 μm

Regole di Scaling dei dispositivi MOSFET

Maggiore integrazione → maggior numero di transistor

Occorre diminuire le dimensioni dei MOSFET

In che modo?

REGOLE DI SCALING

Le regole di scaling determinano il modo in cui i parametri fondamentali del MOSFET devono essere variati per scalare le dimensioni.

Regole di Scaling dei dispositivi MOSFET

Lo scopo è quello di migliorate la

- **densità di integrazione**
- **tempi di ritardo**
- **consumo di potenza**.

Vedremo che non sempre è possibile migliorare significativamente tutti questi parametri in contemporanea

Inoltre, **diminuire L su scale molto piccole (ben al di sotto dei 100 nm)** significa andare incontro a **problematiche di vario tipo**

Scaling a campo costante

Scaling a campo costante

La teoria dello scaling segue in generale tre regole:

1. **Riduzione di tutte le dimensioni verticali e orizzontali di un fattore S**
2. **Riduzione della tensione di soglia e di alimentazione di un fattore S**
3. **Incrementare il livello di drogaggio di un fattore S**

Con S>1

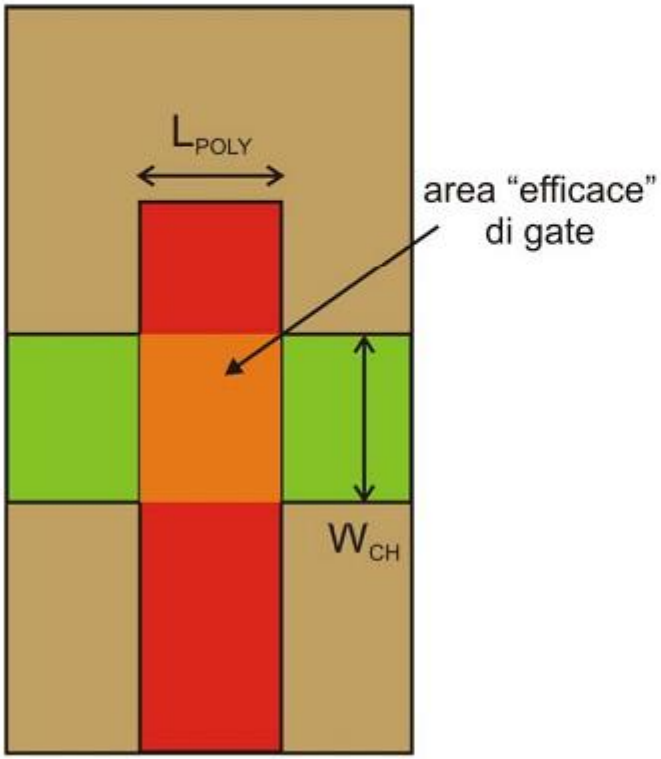
Dato che le dimensioni e il voltaggio scalano di pari passo, tutti i campi rimangono costanti

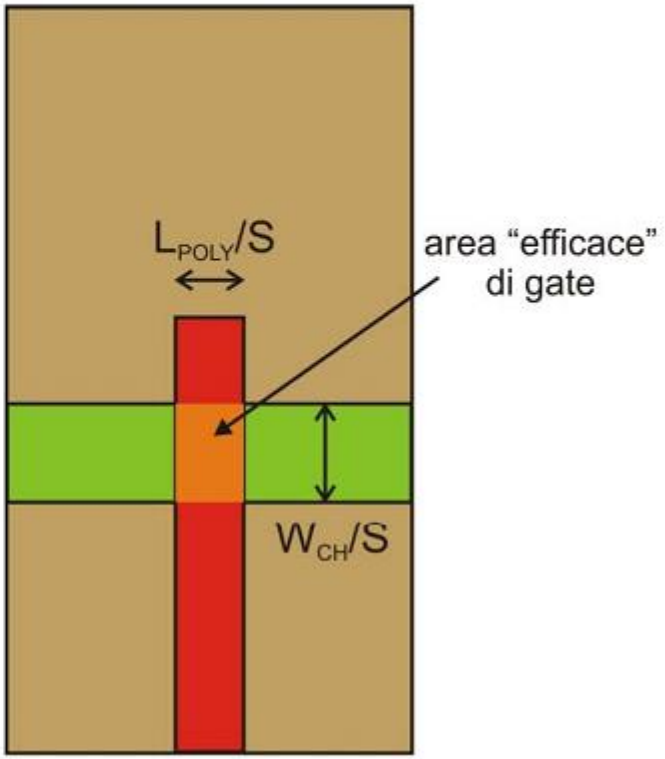
Constant field scaling

Scaling a campo costante

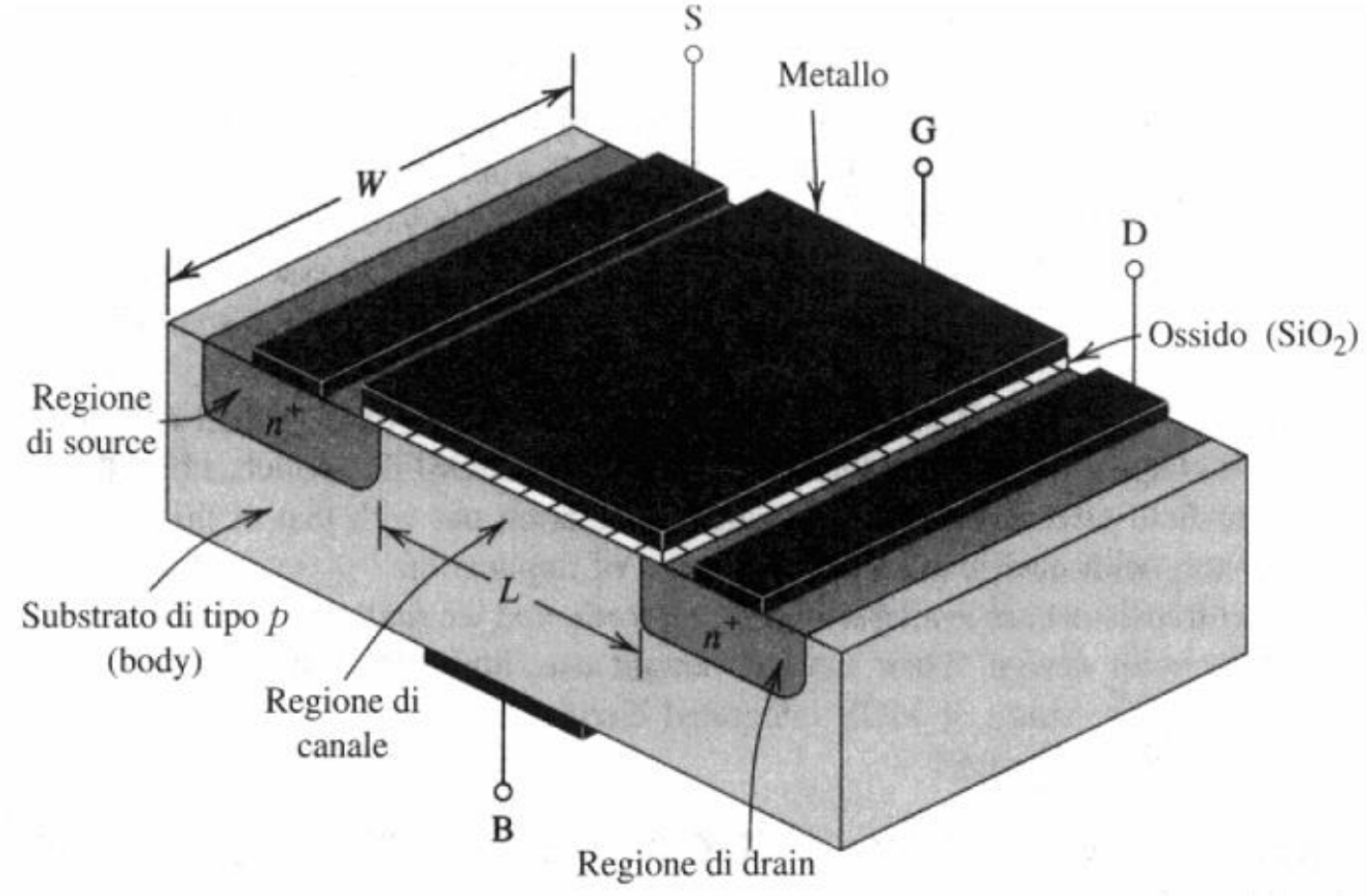
Definiamo le nuove grandezze, scalate in dimensioni di un fattore S, in questo modo:

$$L' = \frac{L}{S} \qquad W' = \frac{W}{S}$$





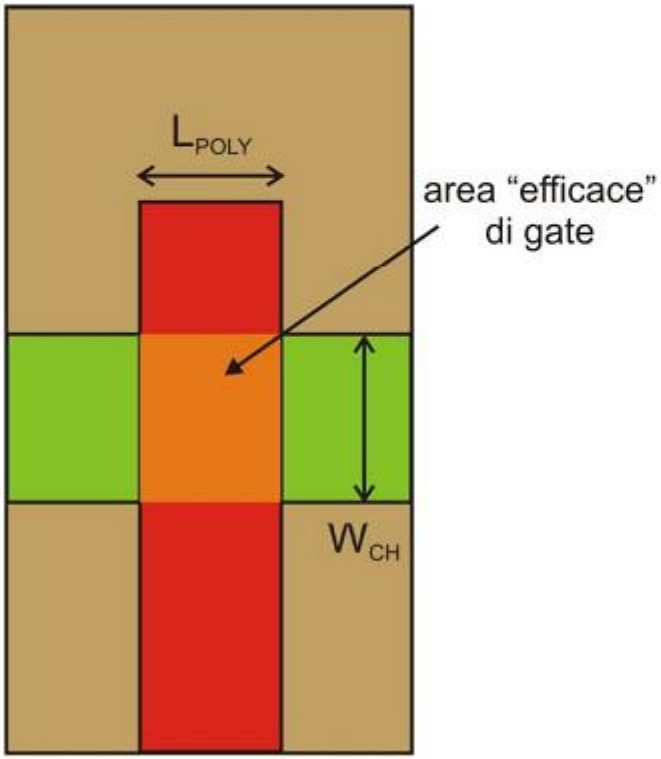
Il sistema MOSFET

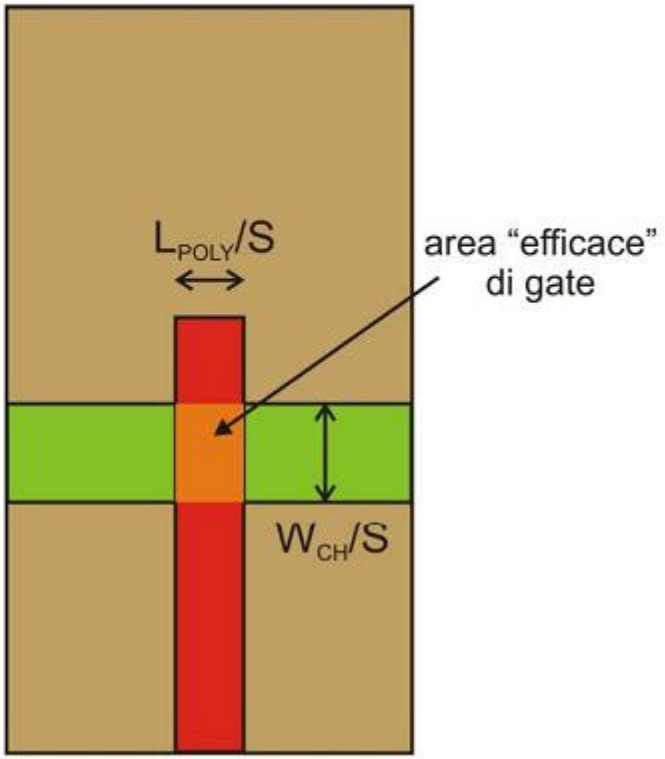


Scaling a campo costante

Definiamo le nuove grandezze, scalate in dimensioni di un fattore S, in questo modo:

$$L' = \frac{L}{S} \qquad W' = \frac{W}{S}$$

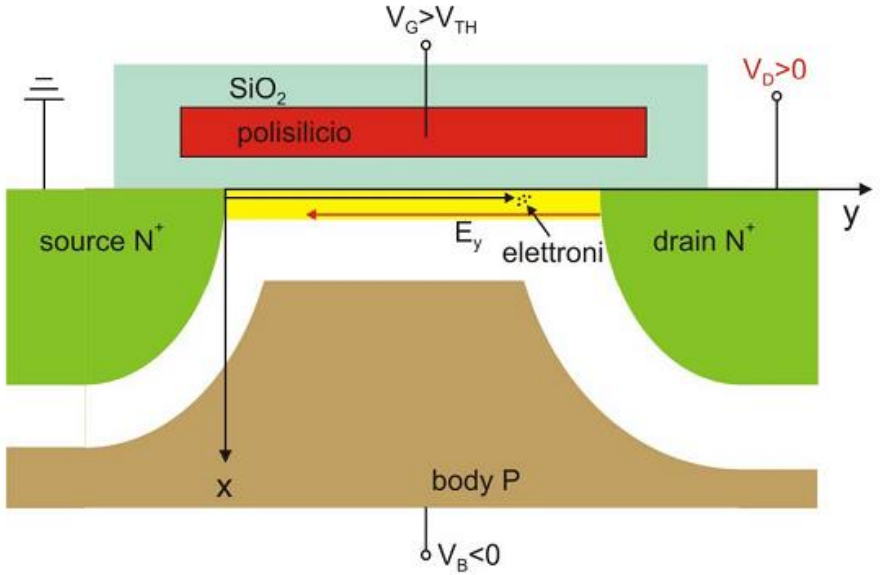




Scaling a campo costante

Nella strategia "a campo costante" **si interviene anche sulla tensione di alimentazione**, e la si scala di un fattore S rispetto al valore della generazione precedente (V_{DD})

$$V_{DD}' = \frac{V_{DD}}{S}$$



Questo comporta che **tutte le tensioni di polarizzazione** in gioco (limitate dalla tensione di V_{DD}) scalino "mediamente" di un fattore S:

$$V_{DS}' = \frac{V_{DS}}{S} \qquad V_{GS}' = \frac{V_{GS}}{S} \qquad V_{SB}' = \frac{V_{SB}}{S}$$

Scaling a campo costante

Cosa succede al campo orizzontale, Ey?

Ci ricordiamo che il canale è formato ($V_{GS}>V_{TH}$) ed applichiamo una tensione tra drain e source $V_{DS}>0$ V.

Il campo elettrico longitudinale in un canale di lunghezza L risulta essere dato approssimativamente da

$$E_y \approx \frac{V_{DS}}{L}$$

$$E_y' \approx \frac{V_{DS}'}{L'} = \frac{V_{DS}}{S}\frac{S}{L} = E_y$$

Lo scalamento non influisce sul campo orizzontale, campo costante!

Scaling a campo costante

Come varia la corrente di drain nell'ipotesi - per semplicità – di essere in regione di pinch-off?

Supponiamo inizialmente di **non eseguire alcuna variazione sulle grandezze tecnologiche** che influenzano la tensione di soglia e che la dipendenza di quest'ultima dalla d.d.p. tra source e body sia trascurabile.

La corrente di drain del dispositivo "scalato" risulta, in tal caso, data da

$$I_D' = K'\left(V_{GS}' - V_{TH}\right)^2$$

Per quel che concerne la transconduttanza, supponendo di non variare lo spessore dell'ossido si ottiene invece:

$$K' = \frac{\mu_N C_{OX} W'}{2L'} = \frac{\mu_N C_{OX} W}{2L} = K$$

Scaling a campo costante

Da cui:

$$I_D' = K\left(\frac{V_{GS}}{S} - V_{TH}\right)^2$$

La corrente di uscita per cui viene ridotta dallo scalamento, a meno che non si decida di intervenire sul processo di fabbricazione con il fine di abbassare la tensione di soglia, ovvero:

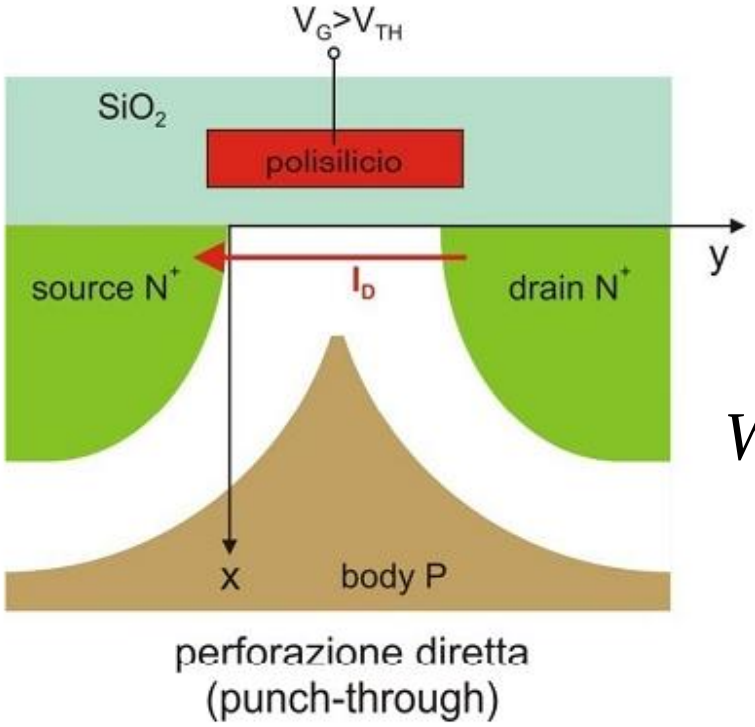
- Livello di drogaggio
- Spessore dell'ossido

$$V_T = \frac{\sqrt{2\varepsilon_S q N_A (2\psi_B)}}{C_o} + 2\psi_B + V_{FB}$$

Scaling a campo costante

Un modo per ridurre la tensione di soglia è sicuramente ridurre il drogaggio degli accettori N_A.

Il canale è corto, per cui **problemi associati al punch-through**



Giunzioni p-n+

W si estende sul lato meno drogato

$$W = \sqrt{\frac{2\varepsilon_s}{q}\left(\frac{N_D + N_A}{N_D N_A}\right)(V_{bi} - V)}$$

$$W = \sqrt{\frac{2\varepsilon_s}{q}\left(\frac{1}{N_D}\right)(V_{bi} - V)}$$

La corrente non è più controllata dal gate!

Non si può diminuire il drogaggio del substrato

Scaling a campo costante

In genere si tende ad aumentare il drogaggio!

Proprio per evitare la perforazione diretta

La tensione di soglia aumenta!

La corrente diminuisce ulteriormente!

$$V_T = \frac{\sqrt{2\varepsilon_S q N_A (2\psi_B + V_{SB})}}{C_o} + 2\psi_B + V_{FB}$$

Inoltre, se il substrato è molto drogato:

- Riduzione tensione di rottura per valanga
- Aumento capacità di svuotamento

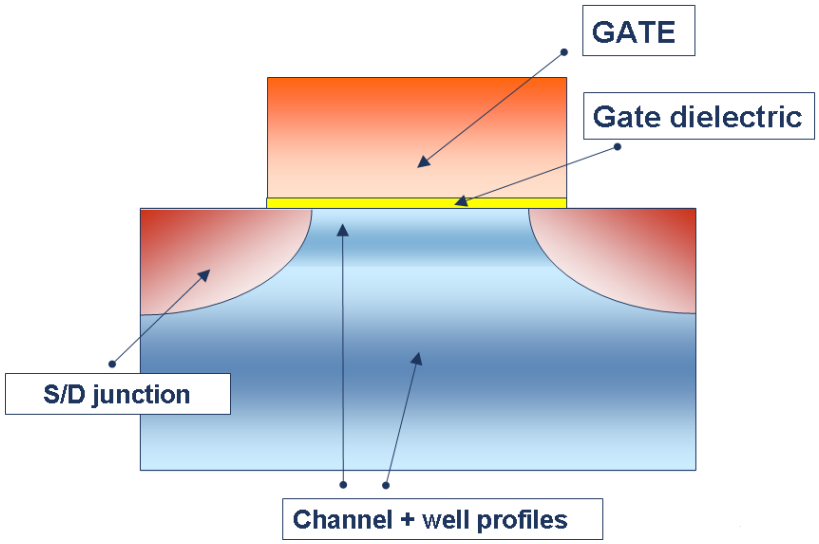
Il drogaggio viene aumentato SOLO nell'area di canale

Impiantazione localizzata

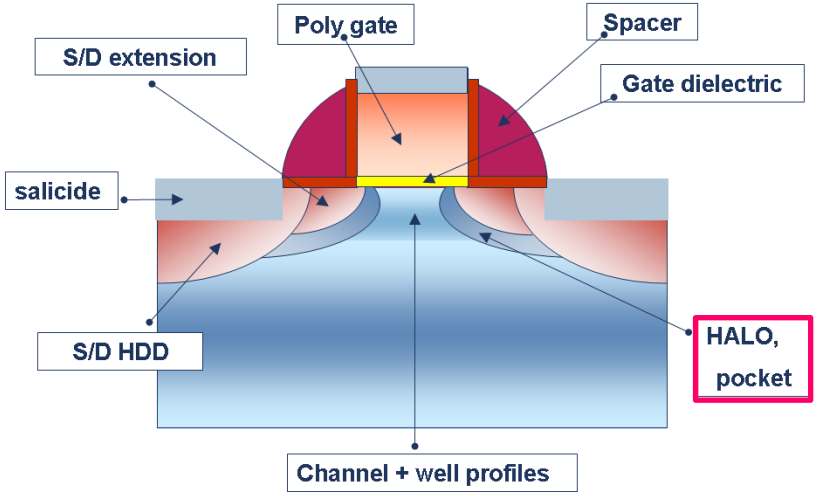
$N_A' = S \times N_A$

Scaling a campo costante

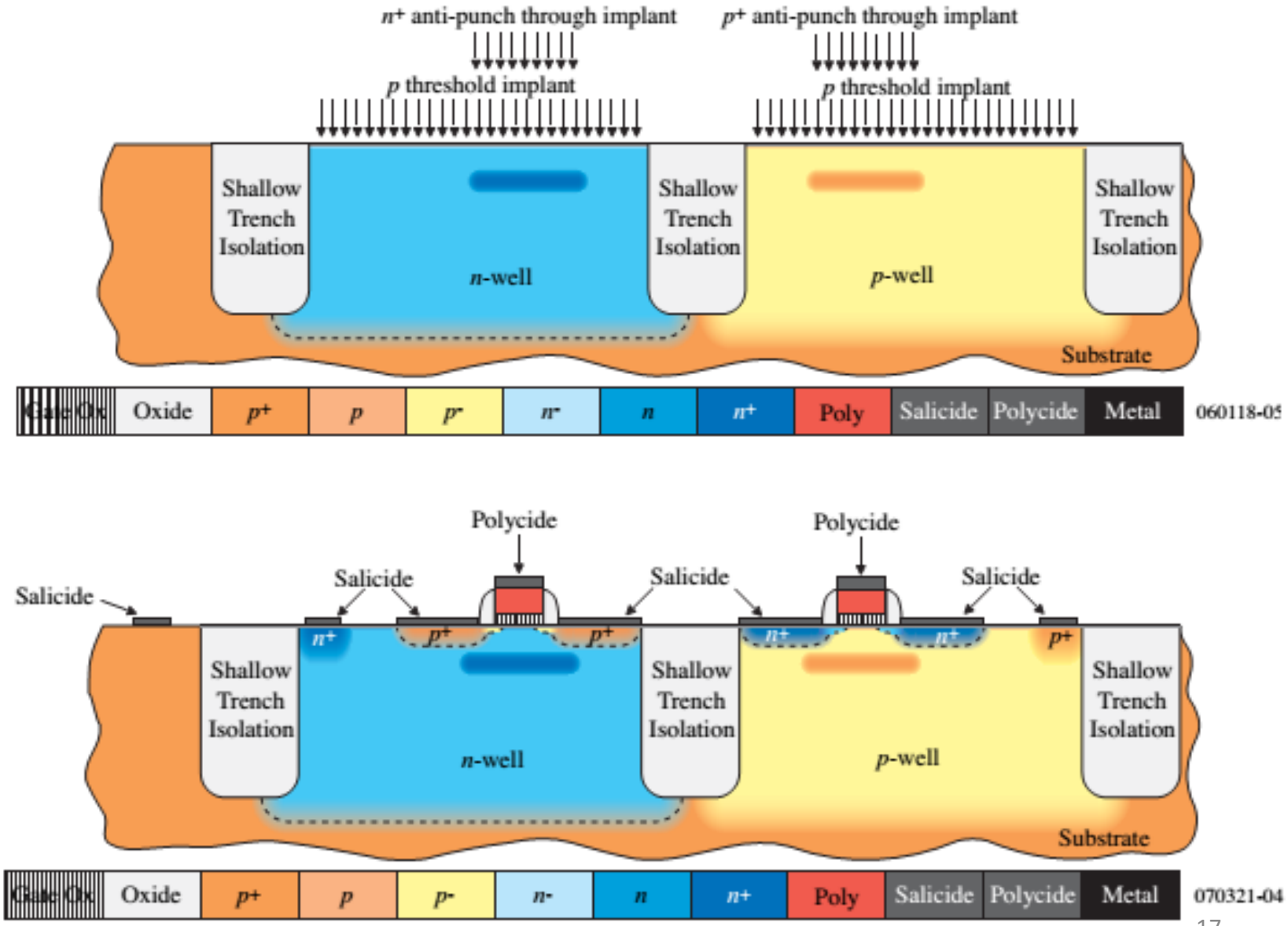
'classical=texbook' MOSFET structure



typical 'advanced' MOSFET Structure



Scaling a campo costante



Scaling a campo costante

Dobbiamo ancora ridurre la tensione di soglia, come faccio?

Scalamento dello spessore dell'ossido

$$d'_{ox} = \frac{d_{ox}}{S}$$

Da cui:

$$K' = \frac{\mu_N C_{OX}' W'}{2L'} = \frac{\mu_N C_{OX}' W}{2L} = \frac{\mu_N \varepsilon_{OX} W}{2t_{OX}' L} = \frac{S \mu_N \varepsilon_{OX} W}{2t_{OX} L} = S \cdot K$$

Lo scalamento ci ha permesso di aumentare K

Cosa succede alla tensione di soglia in virtù di tutte queste modifiche?

È molto complicato dare una risposta immediata!

Scaling a campo costante

Ci ricordiamo che

$$V_{TH} = V_{FB} + 2\psi_B + \gamma\sqrt{2\psi_B + V_{SB}}$$

E che

$$t_{OX}' = \frac{t_{OX}}{S} \quad N_A' = S \cdot N_A \quad V_{SB}' = \frac{V_{SB}}{S}$$

La tensione di soglia sarà maggiormente influenzata dall'effetto body, gli alti due termini variano poco

$$V_{TH} \approx \gamma\sqrt{V_{SB}}$$

$$V_{TH}' \approx \gamma'\sqrt{V_{SB}'}$$

Scaling a campo costante

Ci ricordiamo che

$$\gamma' = \frac{\sqrt{2\varepsilon_S q N_A'}}{C_{OX}'} = \frac{\sqrt{2\varepsilon_S q S N_A}}{\varepsilon_{OX}} \frac{t_{OX}}{S} = \frac{\sqrt{2\varepsilon_S q N_A}}{C_{OX}} \frac{1}{\sqrt{S}} = \frac{\gamma}{\sqrt{S}}$$

Mentre

$$\sqrt{V_{SB}'} = \sqrt{\frac{V_{SB}}{S}} = \frac{\sqrt{V_{SB}}}{\sqrt{S}}$$

Di conseguenza:

$$V_{TH}' \approx \gamma' \sqrt{V_{SB}'} = \frac{\gamma}{\sqrt{S}} \frac{\sqrt{V_{SB}}}{\sqrt{S}} = \frac{\gamma \sqrt{V_{SB}}}{S} = \frac{V_{TH}}{S}$$

Anche la tensione di soglia scala con S

Scaling a campo costante

Come cambia l'espressione della corrente del dispositivo?

$$I_D' = K'\left(V_{GS}' - V_{TH}'\right)^2 = S \cdot K\left(\frac{V_{GS}}{S} - \frac{V_{TH}}{S}\right)^2 = \frac{K}{S}\left(V_{GS} - V_{TH}\right)^2 = \frac{I_D}{S}$$

La corrente viene comunque a diminuire in seguito allo scalamento

Ma questa variazione è meno importante rispetto a quella che abbiamo ottenuto senza variare drogaggio e spessore dell'ossido

Scaling a campo costante

Analizziamo ora le grandezze critiche del sistema, partendo dal campo verticale E_x

Ricordiamo che:

$$V_{DD}' = \frac{V_{DD}}{S} \qquad L' = \frac{L}{S} \qquad W' = \frac{W}{S} \qquad N_A' = S \cdot N_A \qquad t_{OX}' = \frac{t_{OX}}{S}$$

Il campo verticale, in prima approssimazione, vale:

$$E_x \propto E_{OX} = \frac{V_{GS} - 2\psi_B - V(y)}{t_{OX}} \approx \frac{V_{GS}}{t_{OX}}$$

$$E_x' \propto \frac{V_{GS}'}{t_{OX}'} = \frac{V_{GS}}{S}\frac{S}{t_{OX}} = \frac{V_{GS}}{t_{OX}} \Rightarrow E_x = E_x'$$

Lo scalamento ha mantenuto anche il campo verticale costante!

Scalamento a campo costante

Effetti dello scalamento

Supponiamo di considerare un inverter elementare simmetrico

L'area di gate sarà:

$$A_{GATE}' = W_N' L_N' + W_P' L_P' = \frac{W_N L_N}{S^2} + \frac{W_P L_P}{S^2} = \frac{A_{GATE}}{S^2}$$

Occupazione della porta logica minore!

Tempo di propagazione:

$$t_p = \frac{C}{I} V_{DD} = \frac{C}{K} \frac{V_{DD}/2}{(V_{DD} - V_{TH})^2}$$

$$t_P' = \frac{C'}{K'} \frac{V_{DD}'/2}{(V_{DD}' - V_{TH}')^2} = \frac{C}{S} \frac{1}{S \cdot K} \frac{V_{DD}/2S}{\left(\frac{V_{DD}}{S} - \frac{V_{TH}}{S}\right)^2} = \frac{t_P}{S}$$

Tempo di propagazione minore!

La velocità della porta logica aumenta!

Effetti dello scalamento

Frequenza di commutazione:

Immediata conseguenza dell'aumento di velocità è che anche la frequenza di commutazione aumenta di conseguenza di un fattore S

$$f' = S \cdot f$$

Potenza dissipata:

$$P_D' = f' C' \left(V_{DD}'\right)^2 = S \cdot f \frac{C}{S} \left(\frac{V_{DD}}{S}\right)^2 = \frac{P_D}{S^2}$$

Diminuzione notevole della potenza dissipata

- Scalamento della tensione applicata
- Riduzione dell'area e quindi delle capacità parassite

Effetti dello scalamento

Cosa succede invece alla densità di potenza dissipata, ovvero alla potenza dissipata per unità di area?

Ora il numero di porte da integrare è aumentato in funzione dello scalamento del dispositivo!

Potenza dissipata per unità di area

$$p_D' = \frac{P_D'}{A_{GATE}'} = \frac{P_D}{S^2}\frac{S^2}{A_{GATE}} = p_D$$

Riassumendo:

- Lo scalamento **a campo costante** porta ad un miglioramento della maggior parte dei parametri di interesse
- La potenza dissipata per unità di area rimane però costante

Effetti dello scalamento

Precisazioni importanti

Inoltre è inutile scalare le tensioni se lo standard è differente

La maggior parte dei dispositivi vengono comunque alimentati ad una tensione di alimentazione standard!

Inoltre, il ragionamento fatto fino ad ora avrebbe senso se tutte le tensioni scalassero contemporaneamente

Questo per esempio non avviene per le tensioni intrinseche

built in delle giunzioni pn, ampiezza banda proibita → grandezze non scalabili

Inoltre **scalare troppo** la tensione di soglia può far si che diventi **complicato spegnere un MOS**, che può essere acceso in virtù del rumore sulle piste di silicio

Scaling a tensione costante

Alternativa: Scalamento a tensione costante

In questo caso, si preferisce **mantenere la tensione di alimentazione costante**.

Vantaggio dal punto di vista della variazione degli standard

Anche questa tecnica presenta delle problematiche

Le dimensioni del MOSFET verranno comunque scalate (aumentare la possibilità di integrazione)

$$L' = \frac{L}{S} \qquad W' = \frac{W}{S}$$

Ma le tensioni no:

$$V_{DS}' = V_{DS} \qquad V_{GS}' = V_{GS} \qquad V_{SB}' = V_{SB}$$

Alternativa: Scalamento a tensione costante

Il campo orizzontale risulterà amplificato

$$E_y' \approx \frac{V_{DS}'}{L'} = \frac{SV_{DS}}{L} = SE_y$$

In generale quando si opera questa scelta, **si decide comunque di diminuire lo spessore dell'ossido al fine di abbassare la tensione di soglia** (N:B: in tutte e tecniche di scaling si tende a diminuire lo spessore dell'ossido)

$$t_{OX}' = \frac{t_{OX}}{S}$$

Anche in questo caso si opera un **aumento del drogaggio localizzato alla regione di canale**, ma generalmente si aumenta **di un fattore S^2**

$$N_A' = S^2 N_A$$

Scalamento a tensione costante

Cosa succede al campo verticale?

$$E_x' \propto \frac{V_{GS}'}{t_{OX}'} = \frac{SV_{GS}}{t_{OX}} \Rightarrow E_x' = SE_x$$

Di conseguenza il campo verticale **aumenta di un fattore S**

L'aumento dei campi orizzontali e verticali induce però una serie di problematiche:

Rottura dell'ossido di gate

In genere la break down field dell'ossido è intorno ai 6 MV/cm

Scalamento a tensione costante

Cerchiamo ora di valutare l'effetto dello scalamento a tensione costante sugli altri parametri del dispositivo

$$K' = \frac{\mu_N C_{OX}' W'}{2L'} = \frac{\mu_N (S \cdot C_{OX}) W}{2L} = S \cdot K$$

Per quel che concerne la tensione di soglia dobbiamo ricordarci che anche in questo caso vengono comunque effettuate delle modifiche tecnologiche

$$t_{OX}' = \frac{t_{OX}}{S} \qquad N_A' = S^2 N_A$$

Considerando che la tensione di body non viene alterata in seguito alle modifiche, si ottiene:

$$\gamma' = \frac{\sqrt{2\varepsilon_S q N_A'}}{C_{OX}'} = \frac{\sqrt{2\varepsilon_S q S^2 N_A}}{\varepsilon_{OX}} \frac{t_{OX}}{S} = \frac{\sqrt{2\varepsilon_S q N_A}}{C_{OX}} = \gamma$$

$$V_{TH}' \approx \gamma \sqrt{V_{SB}} = V_{TH}$$

Scalamento a tensione costante

Gli effetti della riduzione dello spessore dell'ossido t_{OX} di un fattore S e l'incremento del drogaggio di accettori N_A di un fattore S^2 (dovuto alla criticità della V_D non scalata) si controbilanciano e la tensione di soglia V_{TH} rimane la stessa dei dispositivi della precedente generazione.

$$I_D' = K'\left(V_{GS} - V_{TH}\right)^2 = S \cdot K\left(V_{GS} - V_{TH}\right)^2 = S \cdot I_D$$

A differenza di quanto osservato per lo scaling a campo costante, in questo caso la corrente viene amplificata di un fattore S

Cosa avviene per gli altri parametri significativi?

Scalamento a tensione costante

Area di gate

$$A_{GATE}' = W_N' L_N' + W_P' L_P' = \frac{W_N L_N}{S^2} + \frac{W_P L_P}{S^2} = \frac{A_{GATE}}{S^2}$$

Non abbiamo differenze rispetto al caso precedente, perché in questo parametro non entrano in gioco le tensioni

Tempo di propagazione

La capacità varia allo stesso modo, ma non scalano le tensioni!

$$C' = \frac{C}{S}$$

$$t_P' = \frac{C'}{K'} \frac{V_{DD}/2}{(V_{DD} - V_{TH})^2} = \frac{C}{S} \frac{1}{S \cdot K} \frac{V_{DD}/2}{(V_{DD} - V_{TH})^2} = \frac{t_P}{S^2}$$

Nel caso del full scaling, t_p diminuiva di un fattore S

Scalamento a tensione costante

In generale, se teniamo conto del fatto che

- Le capacità parassite scalano dello stesso fattore
- La corrente di uscita aumenta in questo caso

Con lo scaling a tensione costante si ottengono delle logiche più veloci

Vedremo però che il prezzo da pagare sarà un aumento della potenza dissipata

Scalamento a tensione costante

Frequenza di commutazione

Aumenta maggiormente rispetto al full scaling

$$f' = S^2 \cdot f$$

Potenza dissipata

$$P_D' = f' C' \left(V_{DD}'\right)^2 = S^2 \cdot f \frac{C}{S} V_{DD}^2 = S \cdot P_D$$

Ricordiamo che nel caso del full scaling risultava ridotta di un fattore S^2 rispetto al caso del dispositivo non scalato!

Potenza dissipata per unità di area

$$p_D' = \frac{P_D'}{A_{GATE}'} = S \cdot P_D \frac{S^2}{A_{GATE}} = S^3 \cdot p_D$$

Notevole aumento!

Confronti tecniche di scaling

Parametro	Campo costante (*full scaling*)	Tensione costante
L	1/S	1/S
W	1/S	1/S
t_{OX}	1/S	1/S
C_{OX}	S	S
E_y	1	S
E_x	1	S
K	S	S
V_{DD}, V_{TH}	1/S	1
N_A	S	S^2
I_D	1/S	S

Confronti tecniche di scaling

Parametro	Campo costante (*full scaling*)	Tensione costante
A_{GATE}	$1/S^2$	$1/S^2$
C	$1/S$	$1/S$
t_P	$1/S$	$1/S^2$
f	S	S^2
P_D	$1/S^2$	S
p_D	1	S^3

Full scaling vs tensione costante

Nello **"scaling a campo costante",** la riduzione della tensione di alimentazione V_{DD} **permette di:**

- **evitare incrementi del campo elettrico longitudinale e di quello verticale**
- **di ridurre la potenza dissipata**
- **Tuttavia l'aumento della velocità operativa è limitato** (i circuiti di nuova generazione sono relativamente "lenti")
- la tensione di soglia (ridotta di un fattore S) può diventare troppo bassa e si creano **problemi di compatibilità**

Full scaling vs tensione costante

Per applicazioni in cui **l'aumento di velocità** riveste importanza più elevata rispetto ai consumi, **conviene adottare lo "scaling a tensione costante";**

Questo si paga con:

- **aumento significativo di potenza dissipata** per unità d'area nei circuiti "scalati"
- **eventuali problemi connessi agli elevati campi elettrici** nella regione di canale (elettroni caldi e degradazione della mobilità).

Quindi questa strategia porta a **circuiti veloci, ma con problemi di affidabilità e di consumi elevati.**

Scaling a frequenza costante

Scaling a frequenza costante

Ulteriore strategia, scalamento a frequenza costante

Pensato per **abbattere i consumi**, e per **applicazioni** in cui i circuiti funzionano a delle **frequenze standard** (per esempio per avere batterie più durature perché riduco la dissipazione)

Come al solito, scalo tutte le dimensioni orizzontali e verticali

$$L' = \frac{L}{S} \quad W' = \frac{W}{S} \quad t_{OX}' = \frac{t_{OX}}{S}$$

La principale differenza sta nello scalare la tensione di alimentazione di un fattore S^2

$$V_{DD}' = \frac{V_{DD}}{S^2}$$

La riduzione si riflette su tutte le polarizzazioni del MOSFET

Scaling a frequenza costante

Cosa succede al campo orizzontale?

$$E_y' \approx \frac{V_{DS}'}{L'} = \frac{V_{DS}}{S^2}\frac{S}{L} = \frac{E_y}{S}$$

Il campo si riduce di un fattore S

Per quel che concerne il campo verticale si ottiene:

$$E_x' \propto \frac{V_{GS}'}{t_{OX}'} = \frac{V_{GS}}{S^2}\frac{S}{t_{OX}} = \frac{1}{S}\frac{V_{GS}}{t_{OX}} \Rightarrow E_x' = \frac{E_x}{S}$$

Anche il campo verticale si riduce di un fattore S

Inoltre il fattore K aumenta:

$$K' = \frac{\mu_N C_{OX}' W'}{2L'} = \frac{\mu_N (S C_{OX}) W}{2L} = S \cdot K$$

Scaling a frequenza costante

Ragioniamo ora sulla tensione di soglia

In questo caso **alla riduzione della lunghezza di canale L** corrisponde anche un **forte diminuzione della tensione di polarizzazione**, di un fattore S^2

Di conseguenza il problema dell'estensione della regione di svuotamento è meno importante rispetto ai casi precedenti

Non è necessario aumentare il drogaggio nel canale

$$N_A' = N_A$$

Mentre il fattore di body scala come segue:

$$\gamma' = \frac{\sqrt{2\varepsilon_S q N_A'}}{C_{OX}'} = \frac{\sqrt{2\varepsilon_S q N_A}}{\varepsilon_{OX}} \frac{t_{OX}}{S} = \frac{\sqrt{2\varepsilon_S q N_A}}{S C_{OX}} = \frac{\gamma}{S}$$

Scaling a frequenza costante

In sostanza, la tensione di soglia diventa:

$$V_{TH}' \approx \gamma' \sqrt{V_{SB}'} = \frac{\gamma}{S}\sqrt{\frac{V_{SB}}{S^2}} = \frac{\gamma\sqrt{V_{SB}}}{S^2} = \frac{V_{TH}}{S^2}$$

Pertanto la riduzione significativa della tensione di alimentazione V_{DD} porta ad un crollo (di un fattore S^3) della corrente di drain del transistore MOS

$$I_D' = K'\left(V_{GS}' - V_{TH}'\right)^2 = S \cdot K\left(\frac{V_{GS}}{S^2} - \frac{V_{TH}}{S^2}\right)^2 = \frac{K}{S^3}\left(V_{GS} - V_{TH}\right)^2 = \frac{I_D}{S^3}$$

Vedremo poi che ciò si ripercuote su velocità operativa e potenza dissipata dall'invertitore elementare in tecnologia CMOS.

Scaling a frequenza costante

Area di gate:

Come nei casi precedenti scala di un fattore S^2

$$A_{GATE}' = W_N' L_N' + W_P' L_P' = \frac{W_N L_N}{S^2} + \frac{W_P L_P}{S^2} = \frac{A_{GATE}}{S^2}$$

Tempo di propagazione

$$t_P' = \frac{C'}{K'} \frac{V_{DD}'/2}{(V_{DD}' - V_{TH}')^2} = \frac{C}{S} \frac{1}{S \cdot K} \frac{V_{DD}/2S^2}{\left(\frac{V_{DD}}{S^2} - \frac{V_{TH}}{S^2}\right)^2} = t_P$$

Il tempo di propagazione non varia, per cui non varia la frequenza di commutazione

N.B. questa strategia è utilizzata SOLO per abbassare i consumi

Scaling a frequenza costante

Potenza dissipata:

$$P_D' = f'C'(V_{DD}')^2 = f\frac{C}{S}\frac{V_{DD}^2}{S^4} = \frac{P_D}{S^5}$$

Potenza dissipata per unità di area:

$$p_D' = \frac{P_D'}{A_{GATE}'} = \frac{P_D}{S^5}\frac{S^2}{A_{GATE}} = \frac{p_D}{S^3}$$

Enorme riduzione della dissipazione di potenza, ma non delle prestazioni elettriche

Confronti tecniche di scaling

Parametro	Campo costante (*full scaling*)	Tensione costante	Frequenza costante (*low power*)
L	1/S	1/S	1/S
W	1/S	1/S	1/S
t_{OX}	1/S	1/S	1/S
C_{OX}	S	S	S
E_y	1	S	1/S
E_x	1	S	1/S
K	S	S	S
V_{DD}, V_{TH}	1/S	1	$1/S^2$
N_A	S	S^2	1
I_D	1/S	S	$1/S^3$

Confronti tecniche di scaling

Parametro	Campo costante (*full scaling*)	Tensione costante	Frequenza costante (*low power*)
A_{GATE}	$1/S^2$	$1/S^2$	$1/S^2$
C	$1/S$	$1/S$	$1/S$
t_P	$1/S$	$1/S^2$	1
f	S	S^2	1
P_D	$1/S^2$	S	$1/S^5$
p_D	1	S^3	$1/S^3$

Effetti dello scaling

Effetti di canale corto

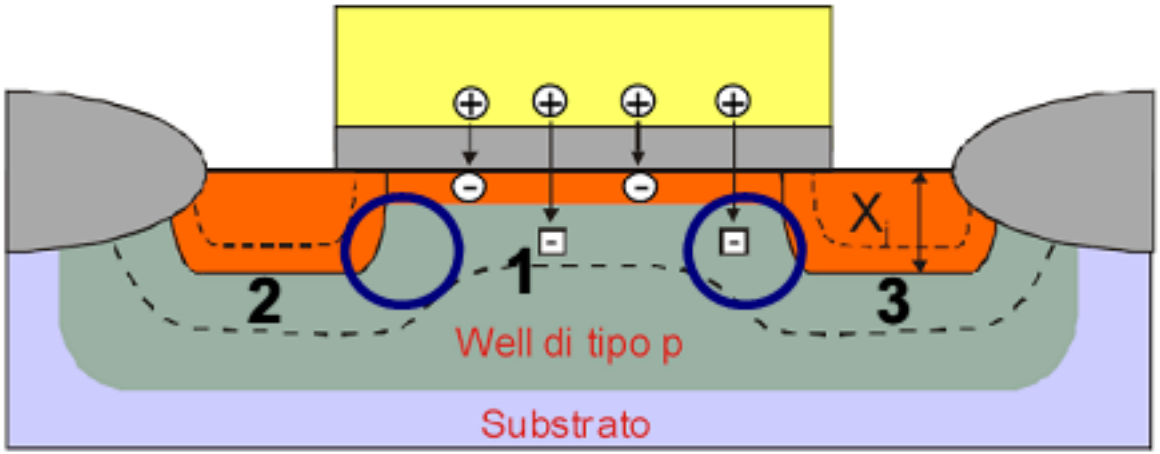
- Variazione della tensione di soglia
- Drain-induced barrier lowering (DIBL)
- Punch Through
- Degradazione della mobilità
- Saturazione della velocità
- Modulazione della lunghezza di canale e variazione dell'impedenza di uscita con la V_{DS}
- Hot carrier effect
- Tunneling attraverso l'ossido di gate

Variazione della tensione di soglia (Vt roll-off)

In una struttura MOS è possibile individuare 3 regioni di svuotamento (si veda la figura seguente):

La regione 1 indotta dall'applicazione della V_G (ioni accettori carichi negativamente perché le lacune sono state "respinte" verso il basso)

le regioni 2 e 3 associate alle due giunzioni P-N body-source (a sinistra) e body-drain (a destra).



Variazione della tensione di soglia (Vt roll-off)

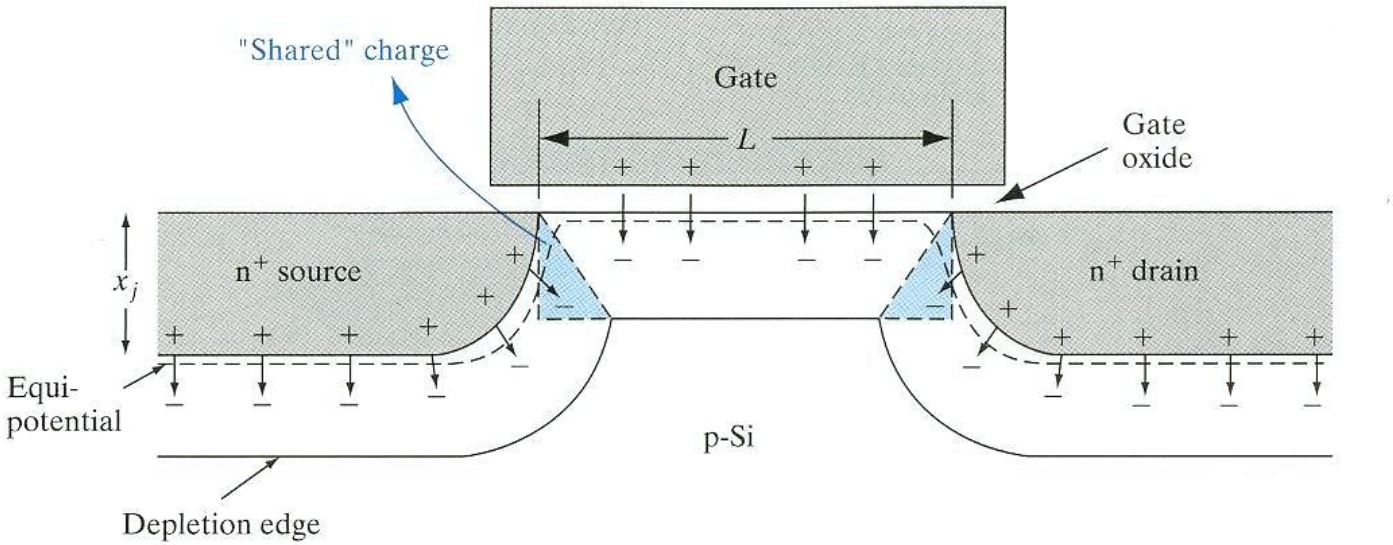
Nella figura è possibile osservare che le regioni 2 e 3 si sovrappongono parzialmente alla regione 1 (cerchi blu).

Pertanto, **le cariche negative presenti nella regione 1 non saranno più tutte "associate" alle cariche positive di gate**, ma anche alla carica positiva delle regioni di svuotamento 2 e 3 (donatori carichi positivamente nel source e nel drain).

Charge sharing

Di conseguenza, si può "intuitivamente" affermare che **parte delle cariche positive di gate risultano "libere"** da vincoli con un determinato numero di cariche negative fisse (N_A^-) della regione 1 e **possono indurre un quantitativo aggiuntivo di elettroni nel canale a parità di V_G applicata.**

Variazione della tensione di soglia (Vt roll-off)

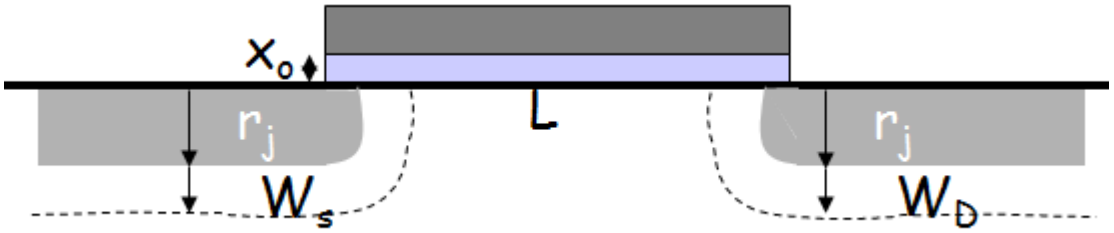


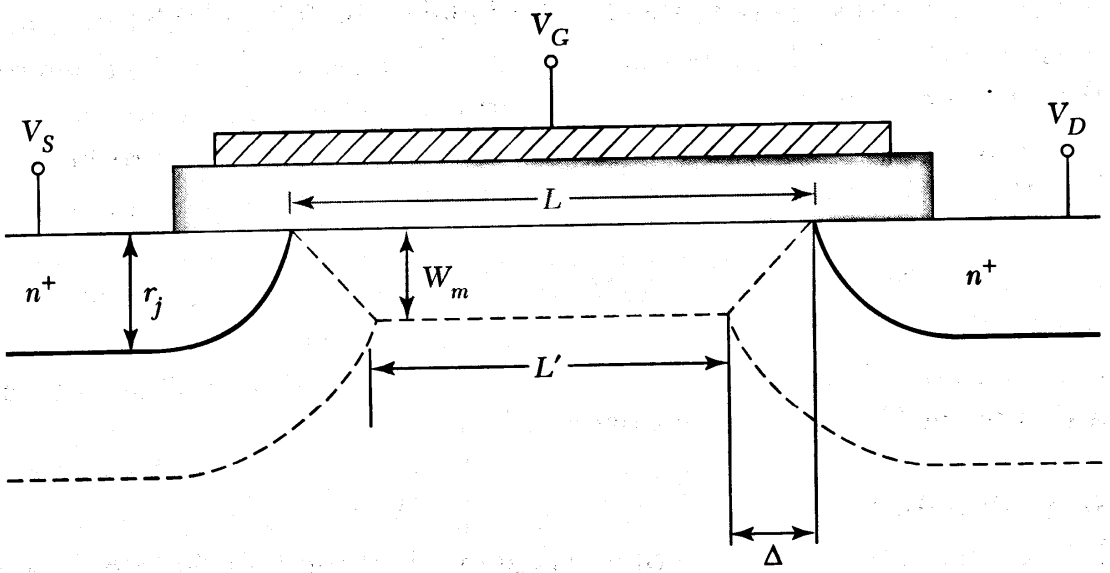
Tali sovrapposizioni risultano trascurabili nel caso di MOS "lunghi" (di vecchia generazione), ma diventano importanti nel caso di MOS "corti".

Questo significa che nel caso di un MOS "a canale corto" la V_G necessaria per indurre una quantità desiderata di carica nel canale è più bassa che nel caso di un MOS "a canale lungo"

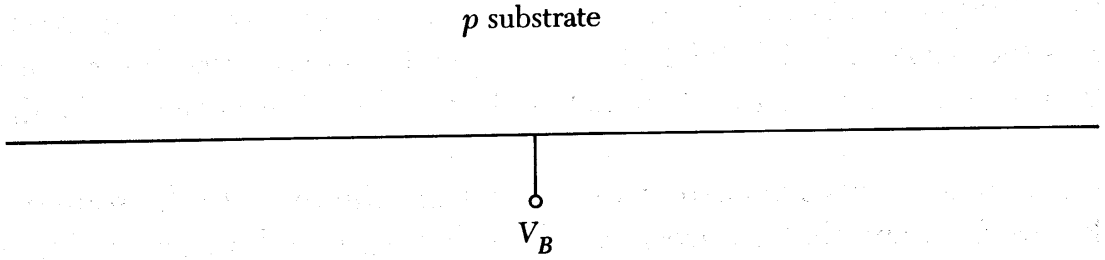
(il che equivale a dire che **la tensione di soglia V_{TH} si riduce al diminuire di L**).

Effetti di canale corto

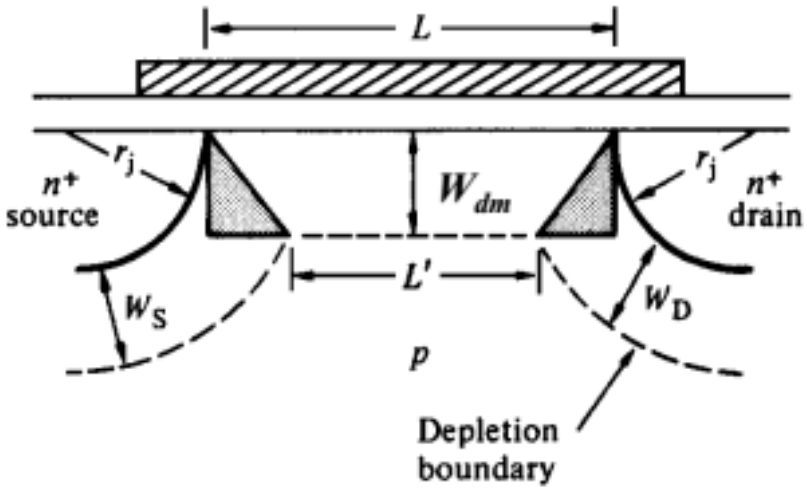




Quanta carica del canale viene controlla dal gate?



Variazione della tensione di soglia (Vt roll-off)



Il gate controlla solo la quantità di carica all'interno di questo trapezio

$$A_{Tr} = \frac{(L + L') \cdot W_{dm}}{2}$$

$$A_r = L \cdot W_{dm}$$

$$\frac{A_{Tr}}{A_r} = \frac{(L + L') \cdot W_{dm}}{2 \cdot L \cdot W_{dm}} = \frac{L + L'}{2L}$$

La carica risulta ridotta di un fattore $1 - \frac{L + L'}{2L}$

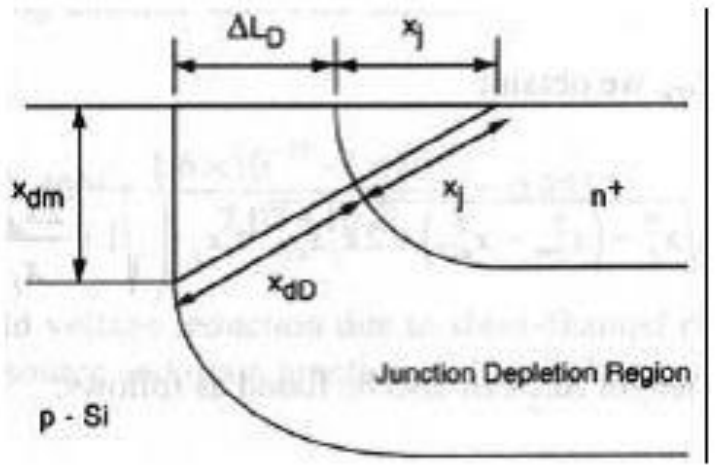
Variazione della tensione di soglia (Vt roll-off)

Consideriamo l'estensione della regione di svuotamento nelle due regioni di Source e di drain

$$x_{dD} = \left(\sqrt{\left(\frac{2\varepsilon_S}{qN_A} \right)(V_{DS} + \phi_0)} \right)$$

$$x_{dD} = \left(\sqrt{\left(\frac{2\varepsilon_S}{qN_A} \right)(\phi_0)} \right)$$

$$\phi_0 = \frac{kT}{q} \ln\left(\frac{N_D N_A}{n_i^2} \right)$$



Si trova che:

$$(x_i + x_{dD})^2 = x_{dm}^2 + (x_j + \Delta L)^2$$

Risolvendo otteniamo:

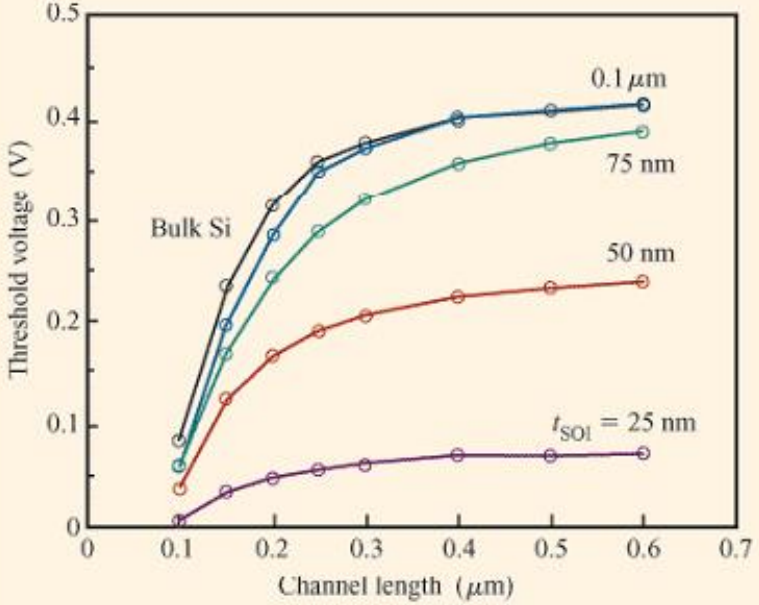
$$\Delta L \cong x_j \sqrt{1 + \frac{2x_{dS}}{x_j}} - 1$$

Variazione della tensione di soglia (Vt roll-off)

Ne segue che la variazione della tensione di soglia sarà:

$$\Delta V_T = -\frac{qN_A x_{dm}}{C_i}\left(1-\frac{L+L'}{2L}\right)$$

$$\Delta V_T = -\frac{qN_A x_{dm}}{C_i}\left(1-\frac{L+L'}{2L}\right) = -\frac{qN_A x_{dm} r_j}{C_i L}\left(\sqrt{1+\frac{2x_{dm}}{r_j}}-1\right)$$

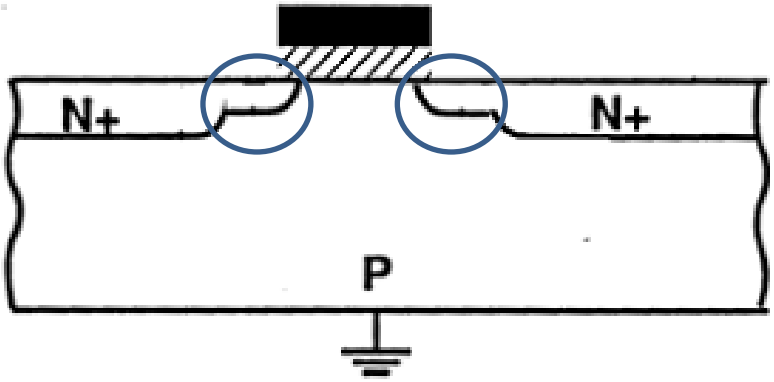


Variazione della tensione di soglia (Vt roll-off)

$$\Delta V_T = -\frac{qN_A W_{max} r_j}{C_i L}\left(\sqrt{1+\frac{2W_{max}}{r_j}}-1\right)$$

Per eliminare, o minimizzare questo fenomeno è necessario:

- Fare delle giunzioni meno profonde (ridurre r_j)
- Aumentare il drogaggio
- Aumentare la Cox, dipende dallo scaling



Questo approccio però aumenta la resistenza di contatto!

$$R_{source}, R_{drain} \propto \rho / W r_j$$

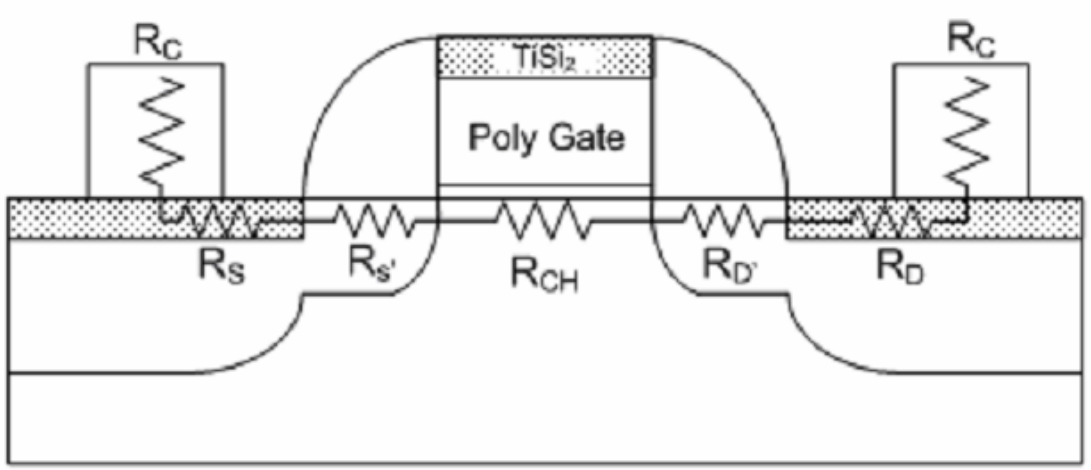
Variazione della tensione di soglia (Vt roll-off)

Sempre con il fine di diminuire l'estensione della regione di svuotamento dei contatti, rispetto a quella del gate, **tali giunzioni vengo fatte con un drogaggio più basso**

Lightly Doped Drain (LDD)

Vedremo successivamente come

Per ridurre ulteriormente la resistenza di contatto viene invece effettuata una deposizione di silicide nell'area dei contatti di source e drain



Drain-Induced Barrier Lowering

Abbiamo già parlato del fenomeno di Charge Sharing dovuto all'estensione delle regioni di svuotamento delle giunzioni p-n dentro il canale.

Se il canale è lungo, la carica del canale è principalmente controllata dal gate, ma al ridursi del canale, abbiamo visto che le cose cambiano.

La regione di svuotamento di Source e Drain possono occupare un parte significativa del canale. Inoltre, all'aumentare di V_{DS}, aumenta la regione di svuotamento al Drain, che quindi può avvicinarsi in maniera significativa al Source.

Se il canale è corto è possibile che source e drain si accoppino

Se questo avviene, il potenziale di iniezione dei portatori al source (che dovrebbe essere controllato dal gate) diminuisce

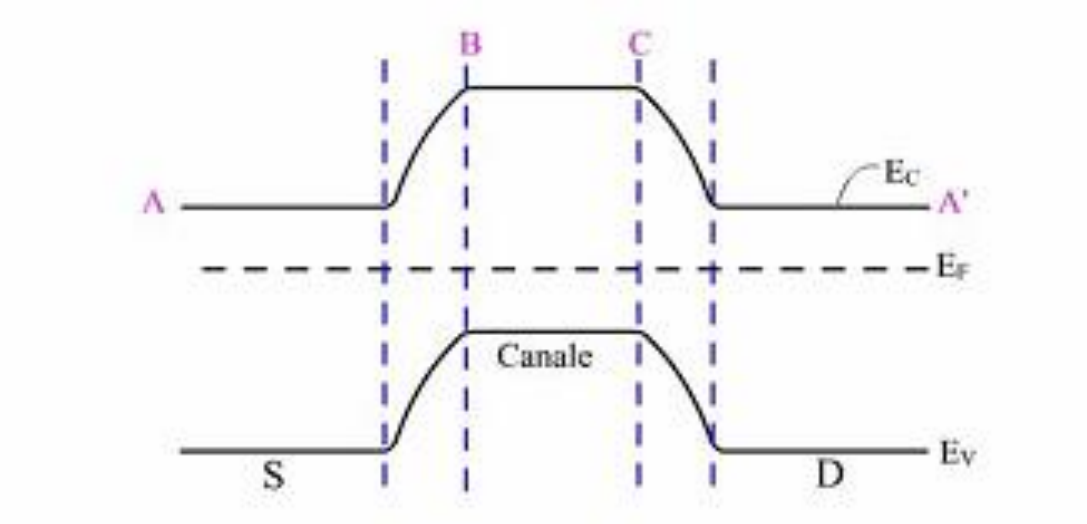
Aumento **significativo della corrente di sottosoglia**

Il sistema MOSFET

Se inizialmente ci poniamo in condizioni di equilibrio, per cui si ha

$V_S=V_{DS}=V_{GS}=V_{BS}=0$

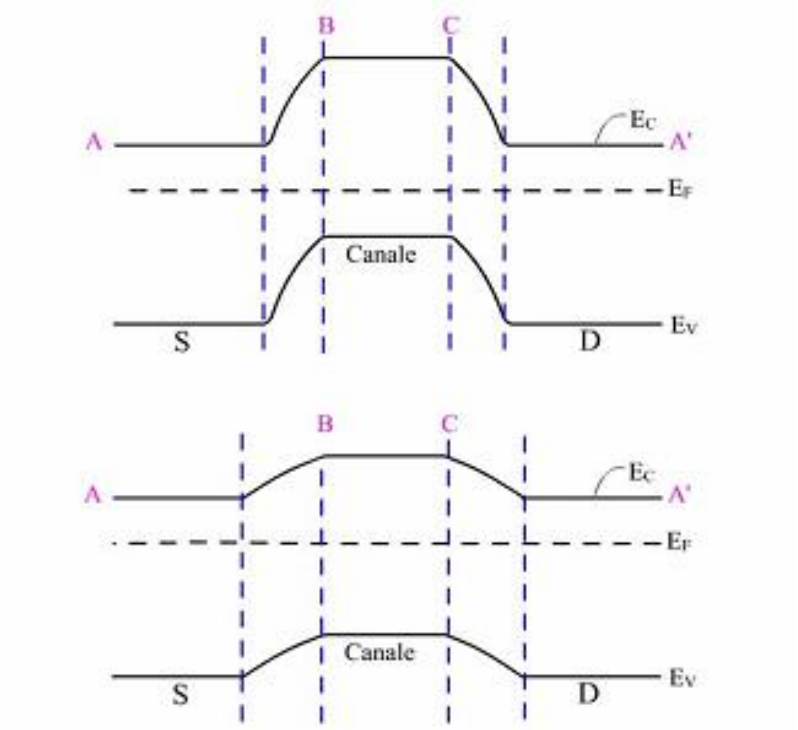
allora avremo una condizione come quella della figura qui sotto, con una distorsione delle bande ottenuta mediante le solite regole delle strutture a bande

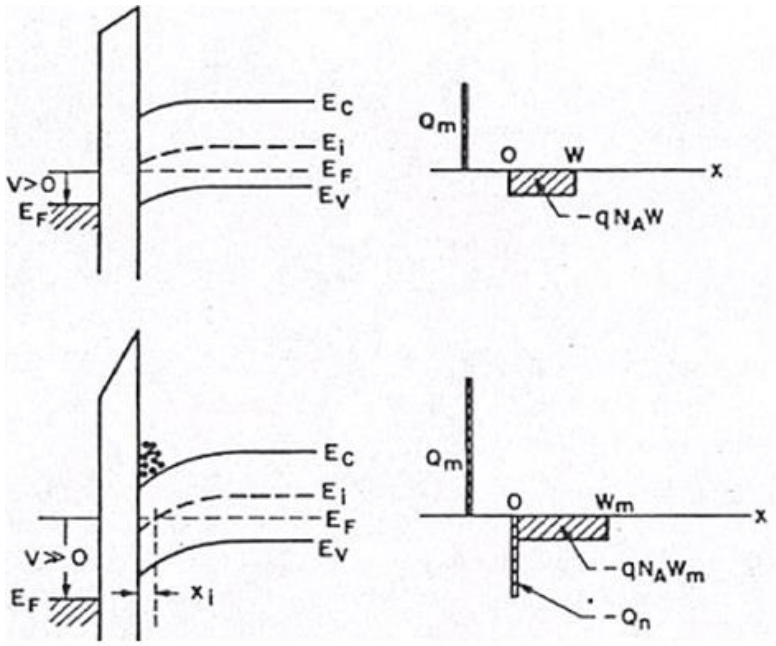


Il sistema MOSFET

Se portiamo il sistema fuori equilibrio ($V_{GS}>0$)

Avremo un'attenuazione della distorsione delle bande per quanto riguarda il substrato, in quanto **vengono richiamati elettroni ed il livello di Fermi dunque sale** (quando arriviamo in inversione, avremo un suo spostamento più in prossimità della banda di conduzione)

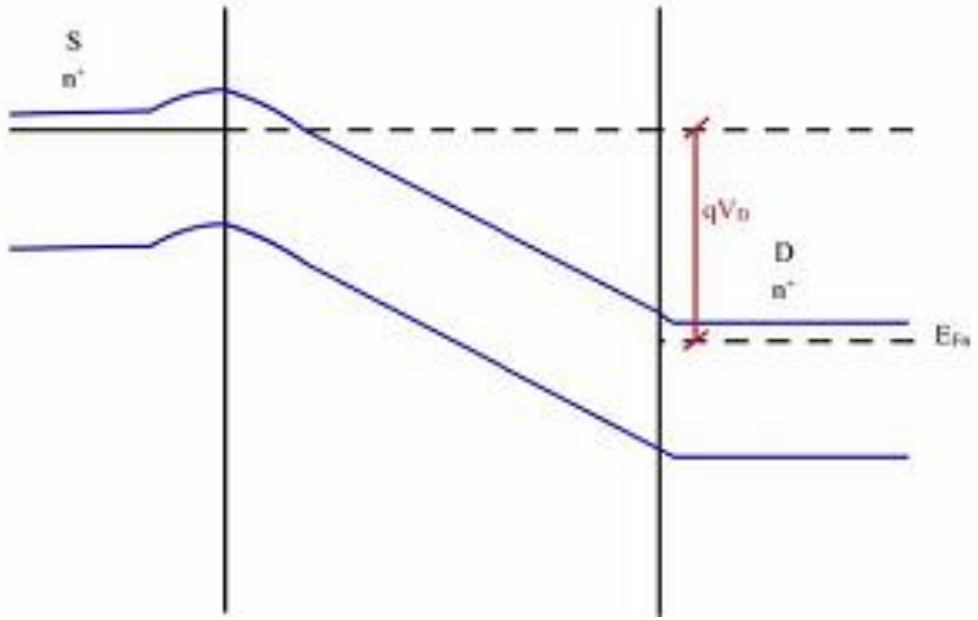




Il sistema MOSFET

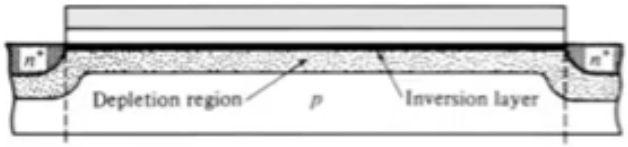
In direzione *y*, invece, la tensione applicata al drain tende a ripartirsi uniformemente attraverso il dispositivo

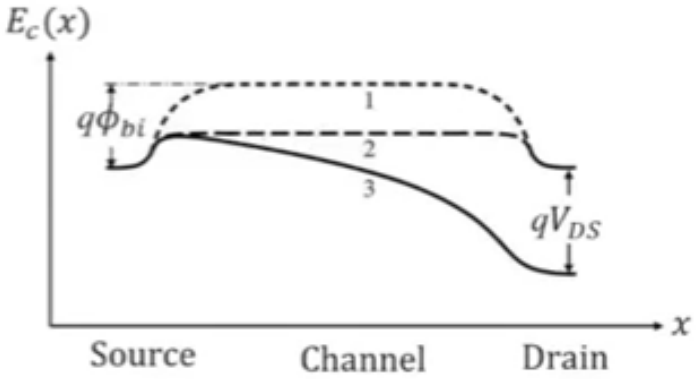
In corrispondenza del source non si avrà alcuna variazione della struttura a bande, mentre, dopo una variazione lineare in corrispondenza del canale, le bande si attesteranno parallelamente ad un **livello di Fermi di drain** che **dista esattamente una quantità pari a qV_{DS} rispetto al** valore individuato nella regione di **source.**



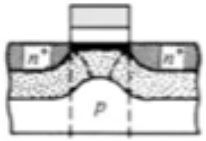
Drain-Induced Barrier Lowering

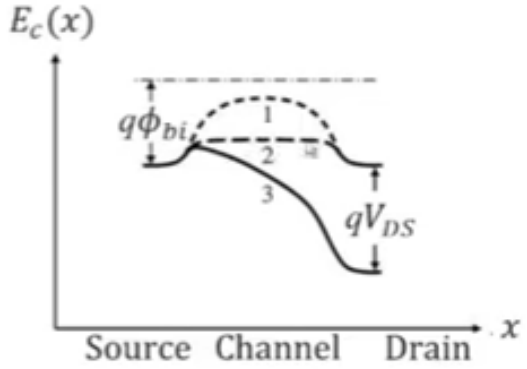
Long channel:





Short channel:

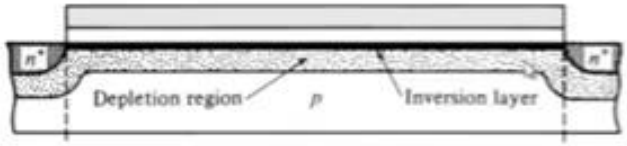


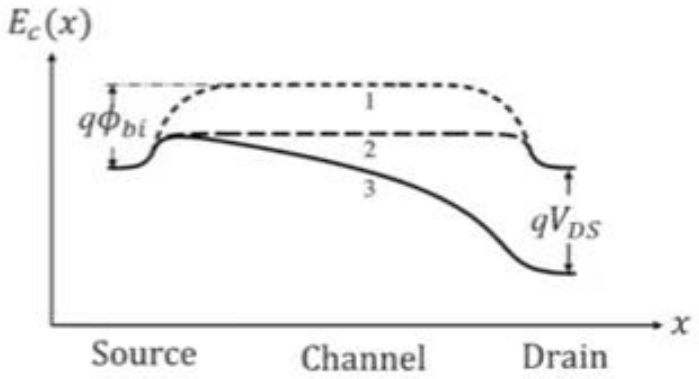


A causa dello charge sharing significativo, il potenziale superficiale del MOS aumenta, in altre parole si abbassa qVbi

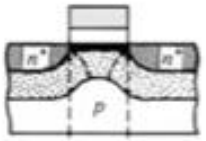
Drain-Induced Barrier Lowering

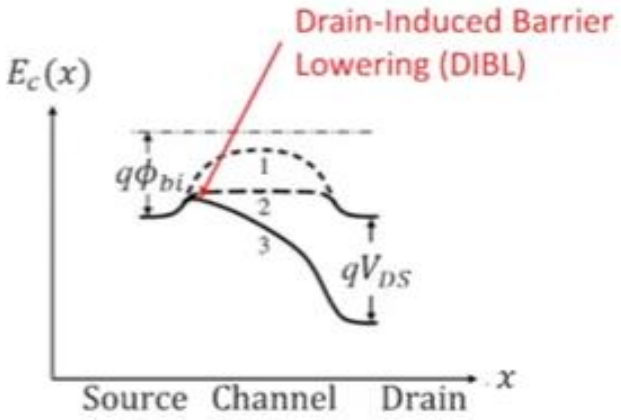
Long channel:





Short channel:



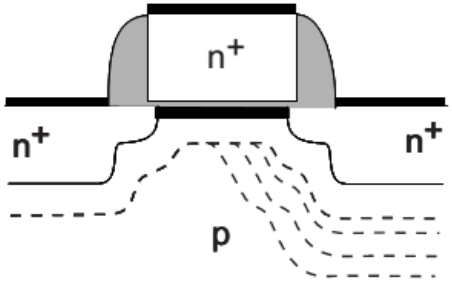


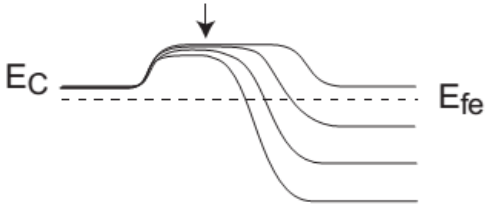
A causa dello charge sharing significativo, il potenziale superficiale del MOS aumenta, in altre parole si abbassa qVbi

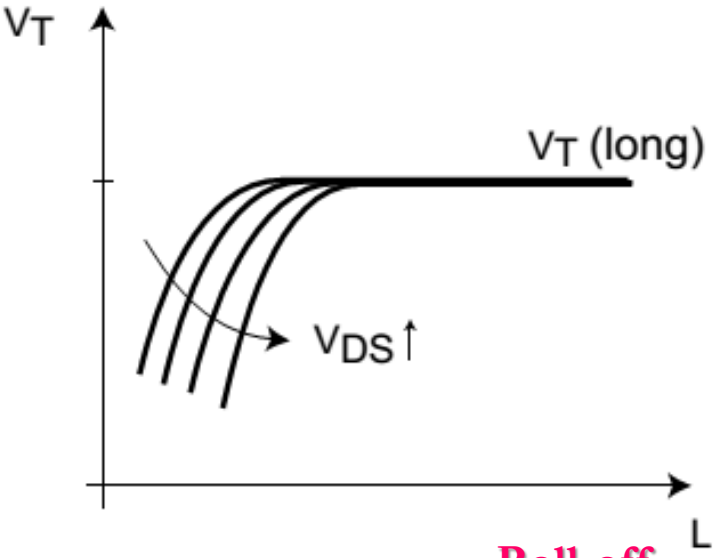
Se il canale è corto, il Source «sente» il potenziale di Drain!

Si abbassa ulteriormente la Barriera all'interfaccia DIBL

Drain-Induced Barrier Lowering

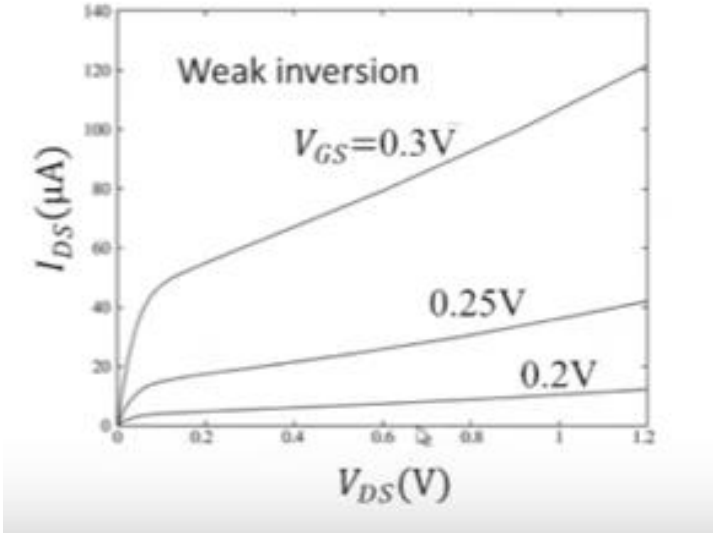






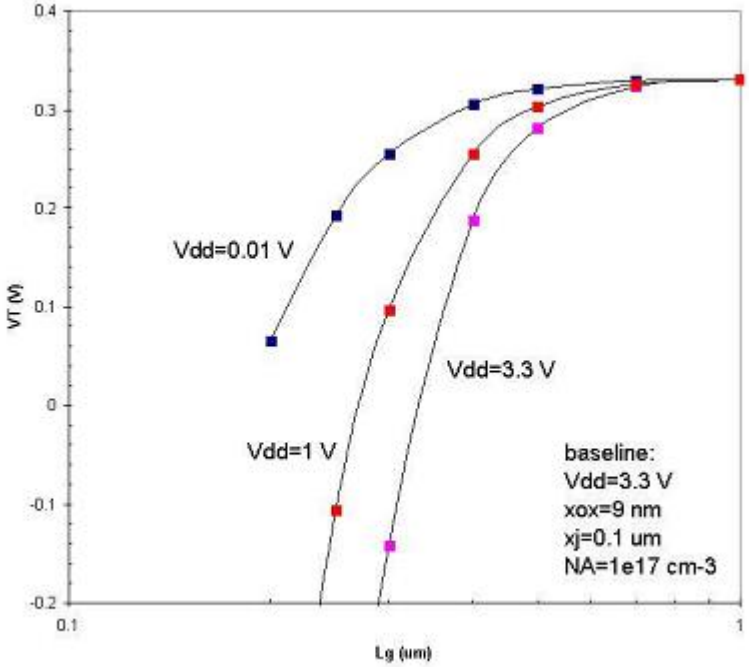
Roll-off

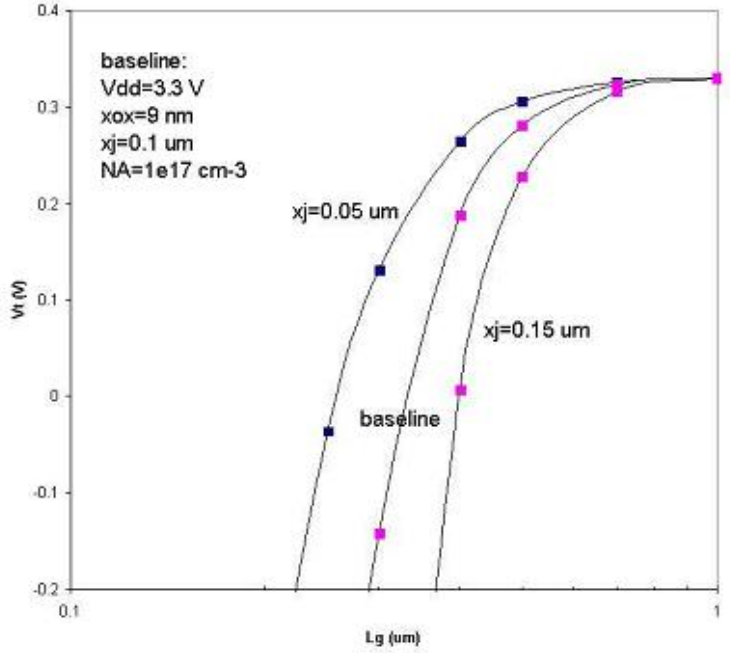
DIBL



Drain-Induced Barrier Lowering

Le regioni LDD, avendo un drogaggio inferiore, riducono il campo e permettono di poter utilizzare tensioni di drain più elevate senza che l'effetto DIBL prenda il sopravvento





Punch-through

Come si è visto, in modo del tutto indipendente dall'approccio adottato, **la lunghezza di canale L viene ridotta di un fattore S**

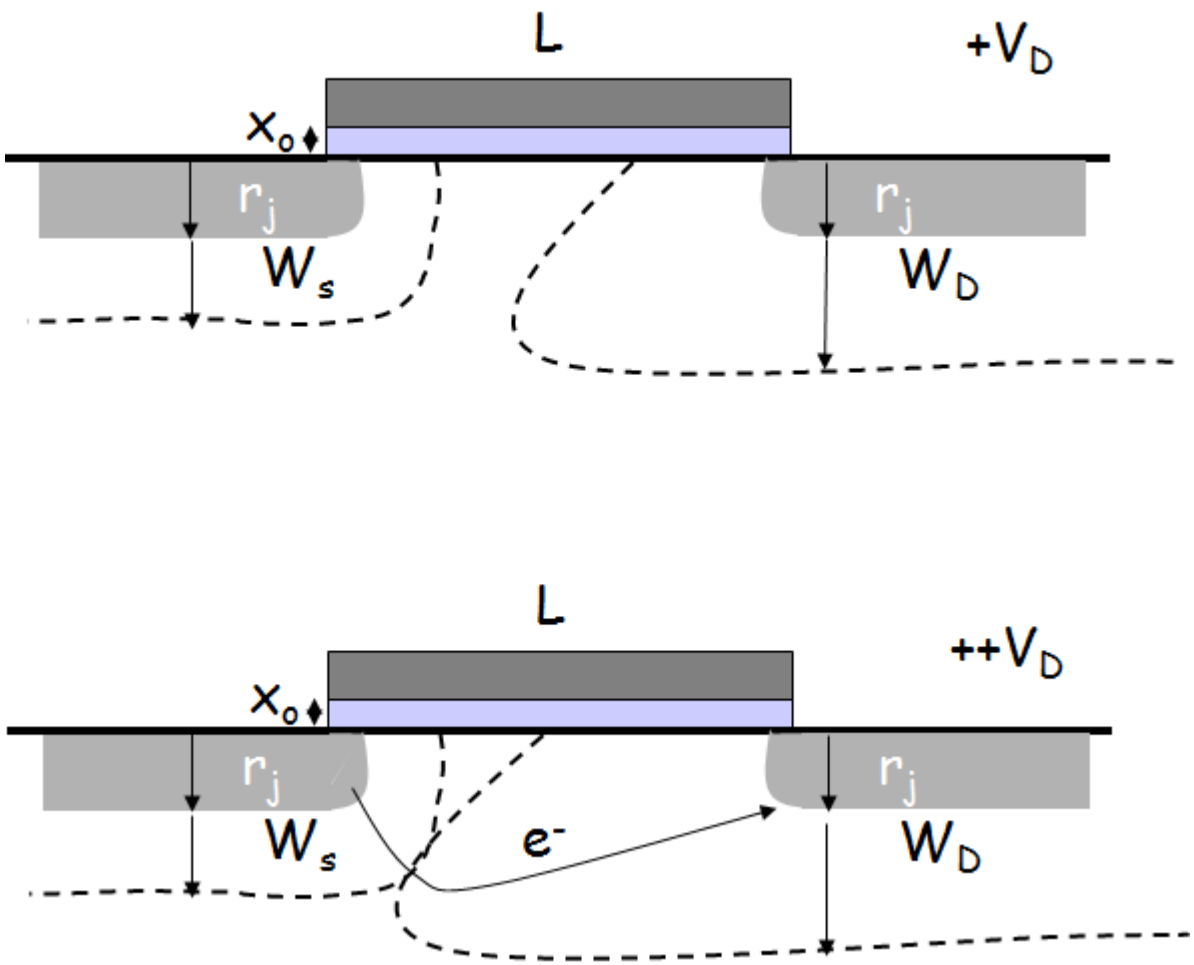
È chiaro allora che, con il procedere della miniaturizzazione, **la possibilità della "perforazione diretta"** (regioni di svuotamento D-B e S-B che si congiungono) **aumenta**

Se si tiene presente che il **drogaggio di body risulta essere molto più basso di quelli di source e di drain**

Allora, **le regioni di svuotamento si estendono principalmente nel body** e sono caratterizzate da uno spessore dato da

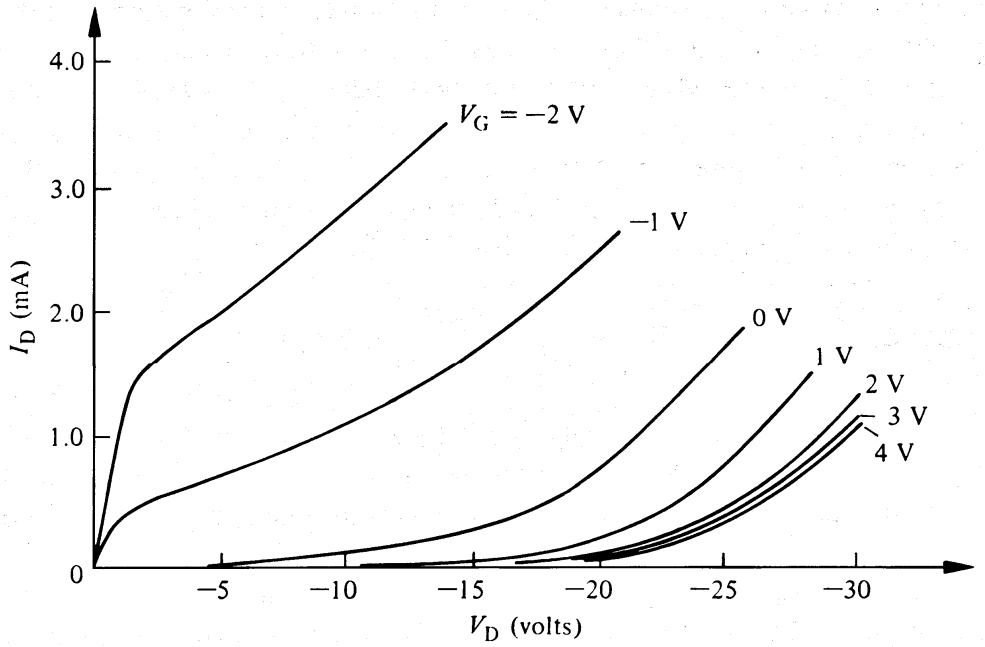
$$d = \sqrt{\frac{2\varepsilon_S \left(V_j + V_R\right)}{qN_A}}$$

Punch-through



Punch-through

- Se avviene il punch-through **non ho più le giunzioni pn back to back**, per cui gli elettroni sono liberi di muoversi da source a drain
- La corrente non è più controllata dal gate
- Il transistor non si spegne



Punch-through

La giunzione che generalmente crea più problemi, ovvero a cui è associata una regione di svuotamento maggiore, è la **giunzione drain-body**

Facciamo un esempio:

Supponiamo che V_{DD}=5 V e teniamo presente che un valore tipico di V_{bi} è 600-900 mV; assumiamo inoltre che V_D sia grossomodo pari al valore della tensione di alimentazione (che rappresenta il worst case per quanto concerne la tensione di contropolarizzazione della giunzione D-B).

In tal caso, nella espressione dello spessore della regione di carica spaziale associata a tale giunzione è possibile trascurare il potenziale di built-in e si ha che:

$$d \approx \sqrt{\frac{2\varepsilon_S V_D}{qN_A}}$$

Proviamo a calcolare quanto varrebbe d per
$N_A = 1x10^{13}\ cm^{-3}$
$N_A = 1x10^{15}\ cm^{-3}$
$N_A = 1x10^{17}\ cm^{-3}$

Punch-through

Scalamento a campo costante

Per evitare la possibilità della perforazione, se L viene scalata di S è evidente che anche d debba essere scalato di S

Tuttavia se si lascia invariata la concentrazione di accettori N_A nella regione di canale, si ottiene che:

$$d' = \sqrt{\frac{2\varepsilon_S V_D'}{qN_A}} = \sqrt{\frac{2\varepsilon_S V_D}{qN_A S}} = \frac{d}{\sqrt{S}}$$

L'estensione della regione di svuotamento scala ma più lentamente

Possibilità di punch-through!!!

Punch-through

Come già detto, i due valori dovrebbero scalare con la stessa velocità.

Per fare questo **occorre aumentare il drogaggio**, localmente.

$$N_A' = S \cdot N_A$$

$$d' = \sqrt{\frac{2\varepsilon_S V_D'}{qN_A'}} = \sqrt{\frac{2\varepsilon_S V_D}{q(S \cdot N_A)S}} = \frac{d}{S}$$

Punch-through

Scalamento a tensione costante

Più probelmatico. Le tensioni non scalano con il dispositivo

Si deve aumentare ulteriormente il drogaggio!

$$N_A' = S^2 N_A$$

$$d' = \sqrt{\frac{2\varepsilon_S V_D'}{qN_A'}} = \sqrt{\frac{2\varepsilon_S V_D}{q\left(S^2 N_A\right)}} = \frac{d}{S}$$

Solo in questo modo è possibile che la regione di svuotamento si riduca di pari passo alla diminuzione della lunghezza di canale

Punch-through

Scalamento a frequenza costante

Molto meno critico!

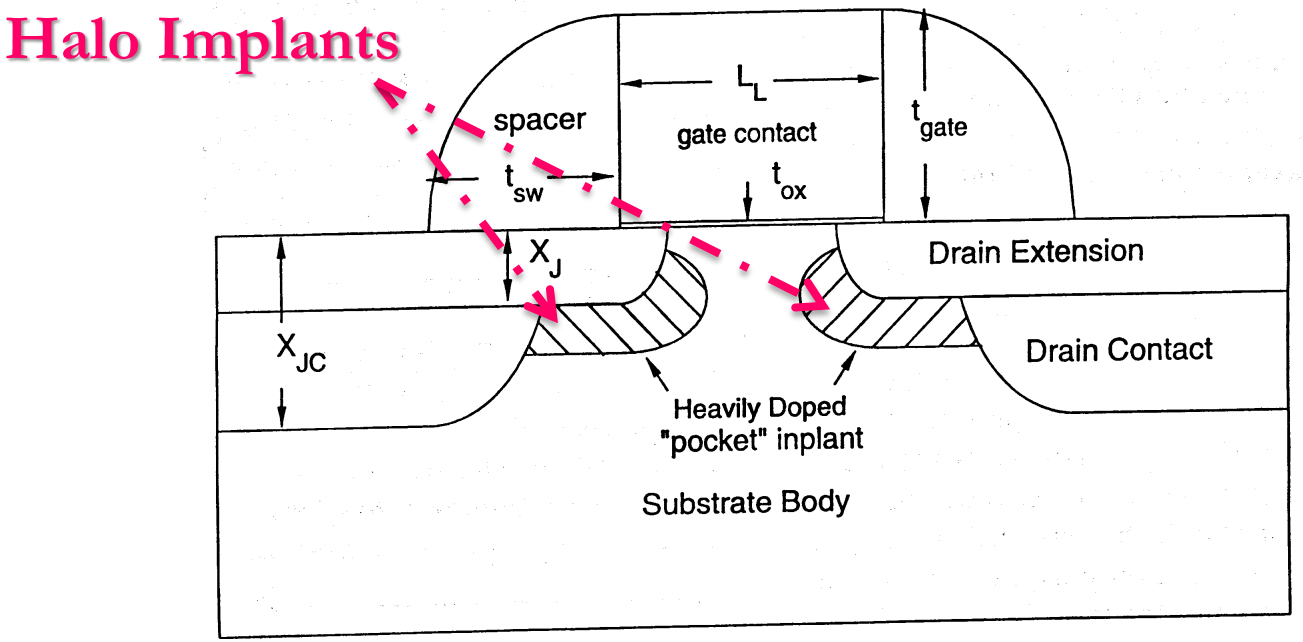
La tensione di polarizzazione scala infatti di un fattore S^2

Di conseguenza l'estensione della regione di svuotamento scalerà di un fattore S, così come la lunghezza di canale

Punch-through

Soluzione

- Drogaggio del substrato più elevato per ridurre le capacità parassite
- Regione ad elevato drogaggio sttostante la regione di canale, per ridurre l'estensione della regione di svuotamento



Degradazione della mobilità

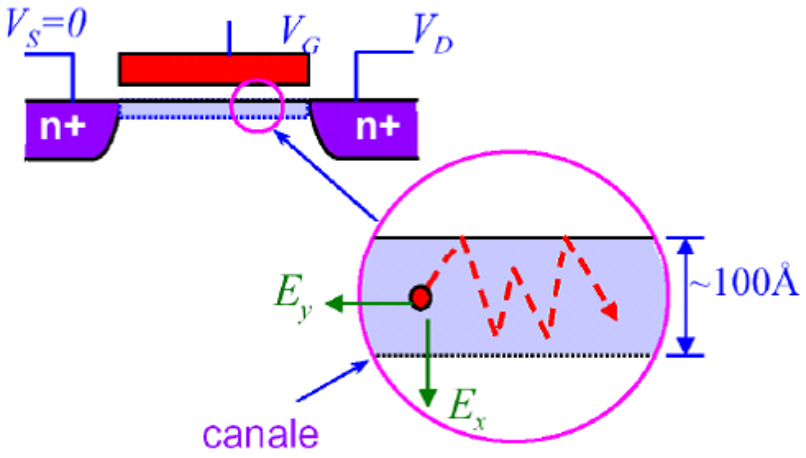
Degradazione della mobilità dovuta alle collisioni superficiali

L'aumento del campo verticale, fa si che il flusso dei portatori non sia più bidimensionale

Aumenta la probabilità di collisione con l'ossido (scattering superficiale) e la mobilità diminuisce

Tende ad essere significativo per $E_x > 6x10^4$ V/cm

A parità di V_{GS} aumenta al diminuire dello spessore dell'ossido



Degradazione della mobilità

L'aumento del campo verticale dovuto allo scaling porta a:

- **Coulomb scattering μ_c:** interazione con impurità ionizzate (per bassi campi e elvato drogaggio)
- **Scattering fononico μ_{ph}:** interazione con le vibrazioni reticolari (per campi intermedi)
- **Rugosità superficiale μr :** rugosità della interfaccia Si-SiO_2 (per campi elevati)

L'effetto del campo verticale viene generalmente descritto in termini di **campo efficace E_{eff}**

Degradazione della mobilità

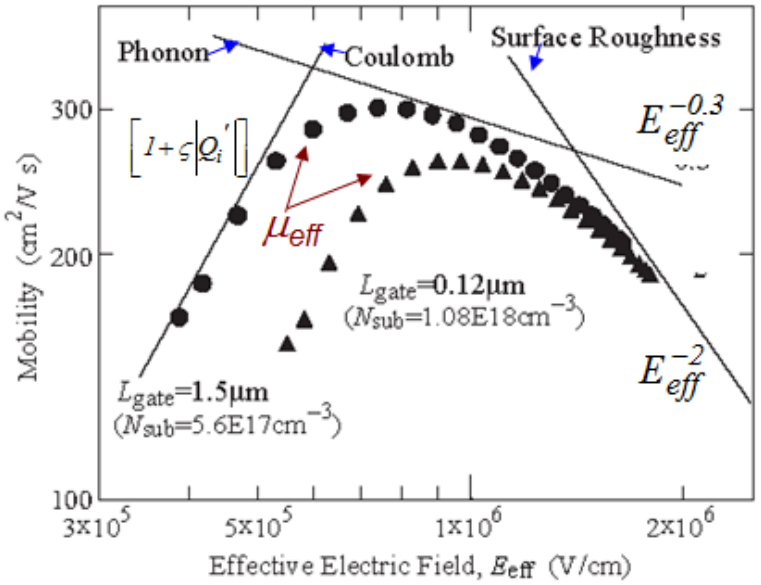
Anche per il campo longitudinale (E_X) la mobilità è descritta da tre contributi

Coulomb scattering $\mu_c \propto \left[1 + \varsigma \left| Q_i' \right|\right]$

Surface scattering $\mu_{sr} \propto \left[E_{eff}\right]^{-2}$

Phonon scattering $\mu_{ph} \propto \left[E_{eff}\right]^{-0.3}$

$$\frac{1}{\mu_{eff}} = \frac{1}{\mu_c} + \frac{1}{\mu_{sr}} + \frac{1}{\mu_{ph}}$$



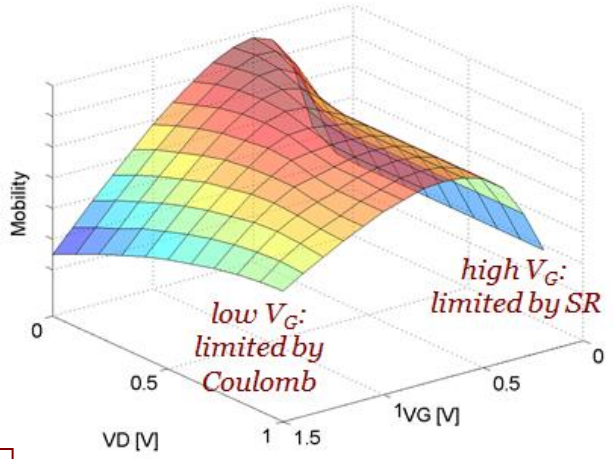
Ad eccezione del primo termine, **in generale la mobilità diminuisce all'aumentare del campo**

Degradazione della mobilità

Integrando le espressioni precedenti si ottiene

$$\int_S^D \frac{I}{\mu(x)} \cdot dx = W \cdot \int_S^D F(Q_i(x)) \cdot dx$$

$$I = \frac{W}{L} \cdot L \cdot \left(\int \frac{1}{\mu(x)} \right)^{-1} \cdot \int_S^D F(Q_i(x)) \cdot dx$$



$$\frac{1}{\mu_{eff}} = \frac{1}{L} \int_S^D \left[\frac{1}{\mu_C(Q(x))} + \frac{1}{\mu_{ph}(Q(x))} + \frac{1}{\mu_{sr}(Q(x))} \right] \cdot dx$$

In sostanza la mobilità dipende dal campo

Possiamo osservare che **esiste un massimo** in cui la mobilità è maggiore

Questa condizione si trova in inversione moderata

Saturazione delle velocità

Saturazione della velocità dei portatori

Inoltre, quando il campo orizzontale supera valori intorno a 10^4 V/cm, la velocità di trascinamento dei portatori tende a saturare.

La velocità non può aumentare al di sopra di un certo limite a causa delle **collisioni dei portatori con gli atomi** che albergano nelle loro posizioni reticolari.

Il limite superiore è detto "velocità di saturazione dei portatori".

Infatti si vede sperimentalmente che, **all'aumentare dell'intensità del campo elettrico longitudinale**, la dipendenza della velocità dei portatori dal campo diviene non lineare, fino a che **la velocità satura.**

Tale fenomeno è **tanto più significativo quanto più il canale è corto** e/o la tensione **V_{DS} è elevata**.

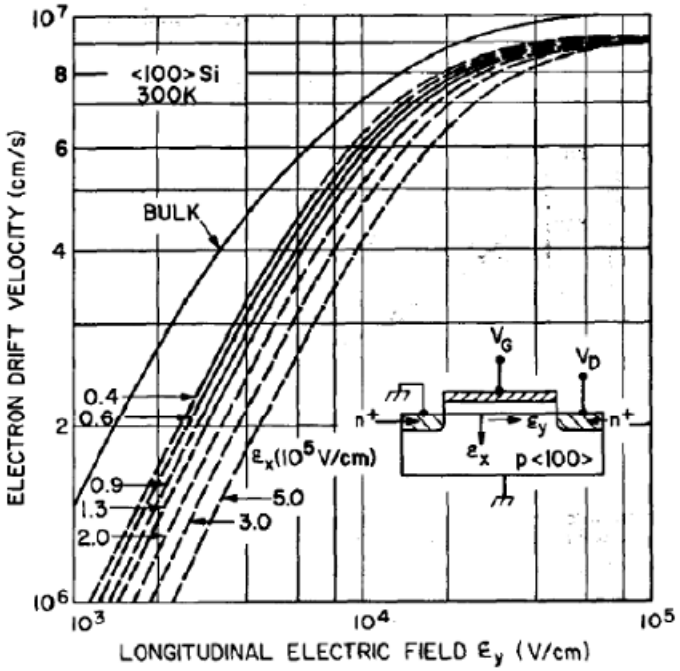
Saturazione delle velocità

Per valori molto elevati del campo, si ha che

$v_N = v_{Nsat}$

Inoltre, **all'aumentare di E_y si ha che μ_N diminuisce**.

Per il silicio si ha che $v_{Nsat}=1\times10^7$ cm/s, mentre $v_{Psat}=6\text{-}8\times10^6$ cm/s



Saturazione delle velocità

Empiricamente si ottiene che

$$v_{drift} = v_{sat} \frac{E_x / E_C}{\left(1 + (E_x / E_C)^{\alpha}\right)^{1/\alpha}}$$

$$\mu \approx \frac{\mu_z}{\left[1 + (E_x / E_C)^{\alpha}\right]^{1/\alpha}}$$

α { 2 for electrons; 1 for holes

In cui $E_C = v_{SAT}/\mu_z$ è il campo critico per il quale la velocità tende a saturare:

Electrons: $v_{sat} \cong 10^5$ m/s $\quad E_c \cong 1$ V/μm
Holes: $v_{sat} \cong 8.10^4$ m/s $\quad E_c \cong 3$ V/μm

Notare che μ_z include anche la riduzione della mobilità dovuta al campo verticale

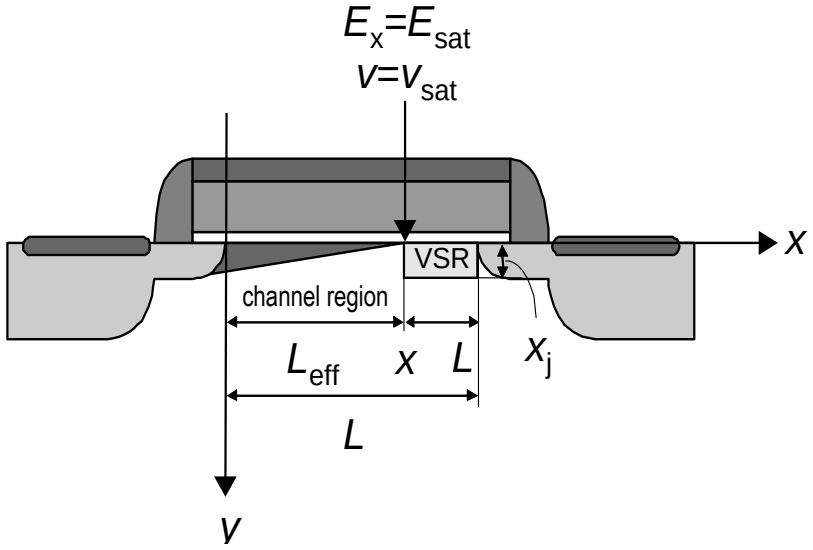
Modulazione della lunghezza di canale

In condizioni di **forte inversione e di saturazione** (Ex>>Ec, saturazione), **la regione di carica spaziale al drain è funzione della stessa V_{DS}** e, di conseguenza, **anche la lunghezza L** del canale ne è funzione (L decresce all'aumentare della V_{DS} applicata).

Il punto di pinch-off si sposta verso il source e il fenomeno diventa importante più il canale iniziale è piccolo

Poiché **la corrente di deriva è inversamente proporzionale alla lunghezza L**, si osserva un incremento di I_{DS} in funzione di V_{DS}

In più ci sarà una regione in cui avviene la velocity saturation



Modulazione lunghezza di canale

Si può dimostrare che

$$\Delta L \cong L_C \cdot \ln\left(\left[\frac{V_{DS}-V_{DSsat}}{L_C \cdot E_C}\right]+\sqrt{1+\left[\frac{V_{DS}-V_{DSsat}}{L_C \cdot E_C}\right]^2}\right)$$

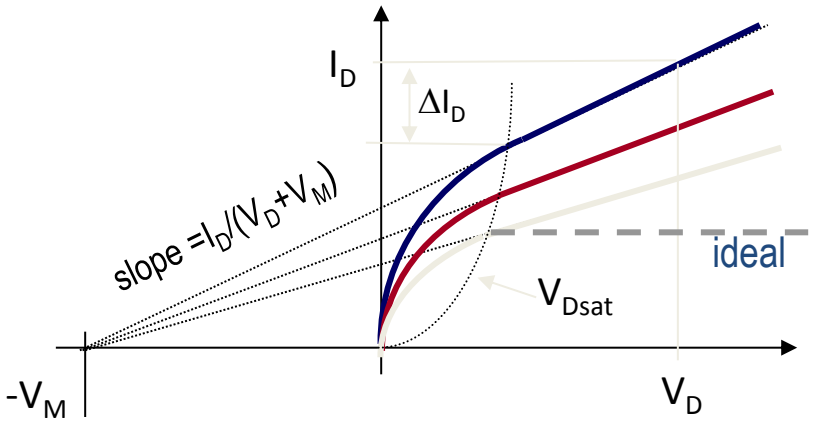
$$L_C = \sqrt{\frac{\varepsilon_{si} \cdot X_J}{C_{OX}'}}$$

junction depth

The smaller the junction depth and oxide thickness, the smaller CLM effect.

La conduttanza di canale diventa

$$g_{ds} = \frac{I_D}{V_M + V_D} = \frac{\Delta I_D}{V_D - V_{Dsat}}$$



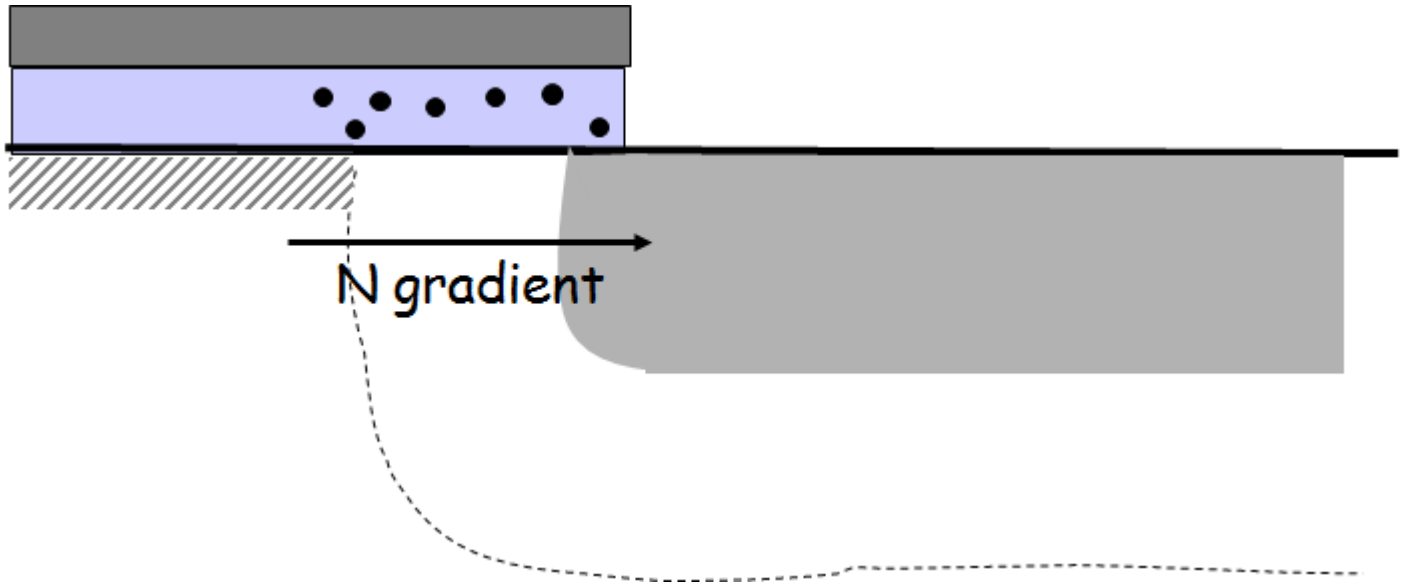
$$\frac{\Delta L}{L} = \lambda V_{DS}$$

$$I_{Dsat} = \frac{Z\mu_n}{2L} C_{ox}(V_G - V_T)^2(1+\lambda V_{DS})$$

Effetto di elettroni caldi

La presenza di campi elevati nella regione di canale fa sì che gli **elettroni siano dotati di energia cinetica elevata** (chiamati appunto portatori "caldi").

Questi elettroni, sotto l'azione congiunta dell'elevato campo orizzontale e del campo verticale, **possono essere addirittura in grado di entrare nell'ossido per tunneling**, a fronte della elevata capacità isolante (dielettrica) garantita da quest'ultimo



Effetto di elettroni caldi

Questo può portare a:

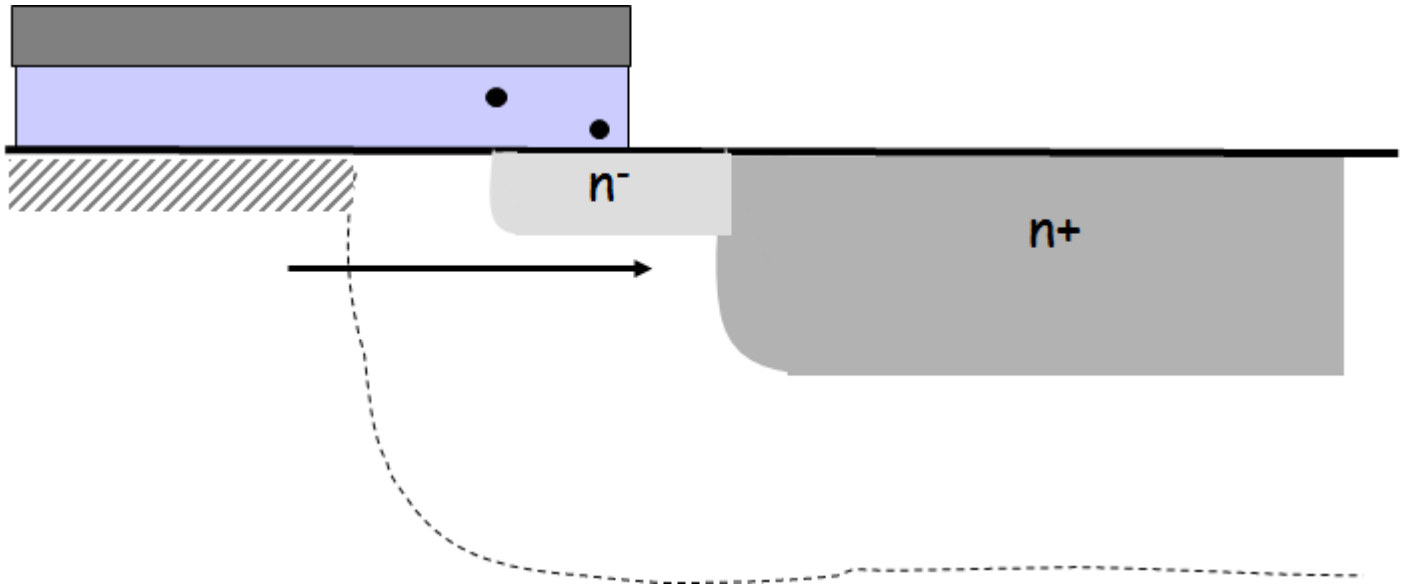
- una **variazione incontrollata della V_{TH}**, che dipende dalla carica nell'ossido attraverso la "tensione di banda piatta" V_{FB};
- una **possibile rottura "a lungo termine" dell'ossido**, la cui qualità degrada nel tempo all'aumentare dell'iniezione di elettroni caldi.

  Questo fenomeno di rottura prende il nome di **time dependent destructive breakdown** (TDDB).

Effetto di elettroni caldi

Ancora una volta, l'utilizzo di regioni meno drogate favorisce una riduzione del campo elettrico

Riduzione del numero di elettroni caldi che possono entrare nell'ossido



Corrente di gate - tunnelling

Lo scalamento del dispositivo fa si che lo **spessore dell'ossido possa diventare molto piccolo** e possa essere ridotto a valori intorno a 1.2 nm.

Per valori così piccoli può essere **possibil**e che una corrente scorra attraverso l'ossido per **effetto tunnel.**

Il fenomeno dipende dalla tensione di gate applicata

Può avere diverse componenti:

- Gate to 'channel' tunneling current (intrinsic)
- Gate to source/drain overlap tunneling current (extrinsic), aumenta nei MOSFET scalati