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

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

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

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 di canale corto

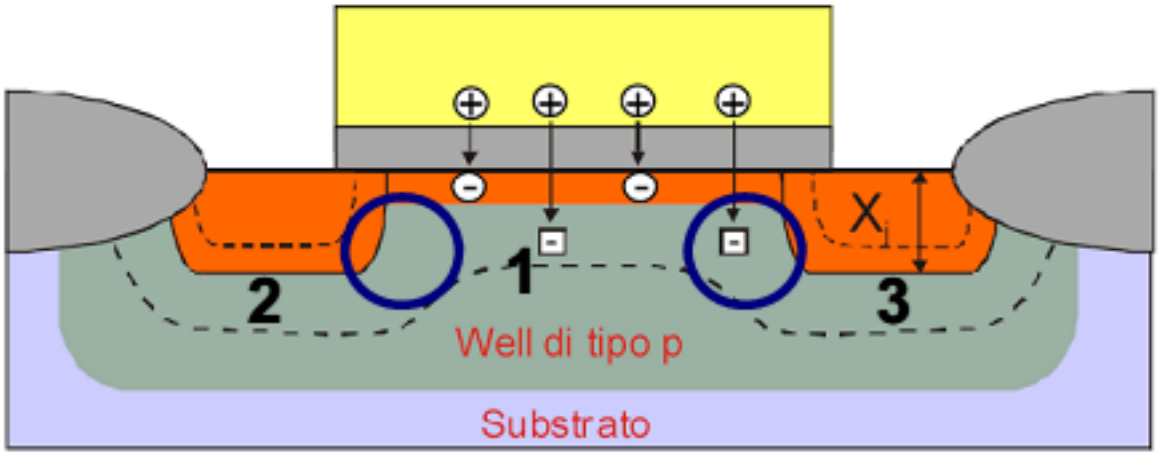
- Variazione della tensione di soglia
- Drain-induced barrier lowering (DIBL)
- Degradazione della mobilità
- Saturazione della velocità
- Modulazione della lunghezza di canale
- Hot carrier effect
- Variazione dell'impedenza di uscita con la V_{DS}

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).



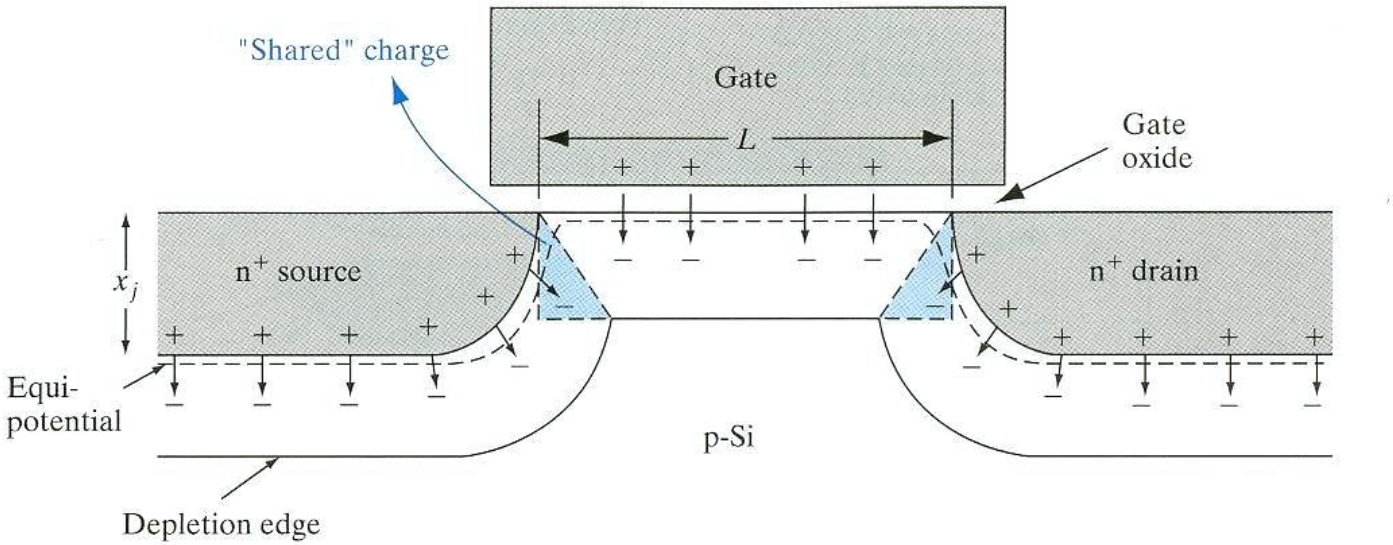
Variazione della tensione di soglia (Vt roll-off)

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).

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)

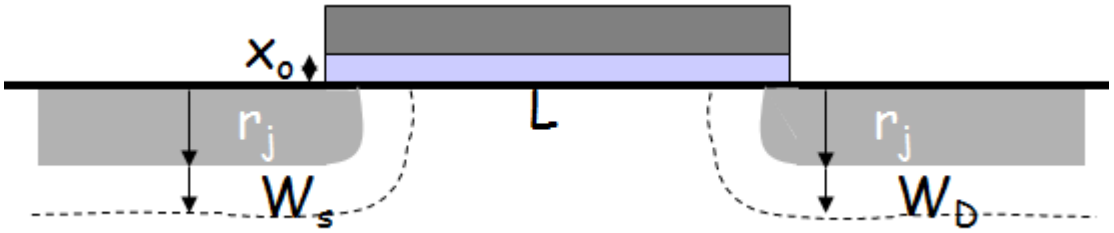


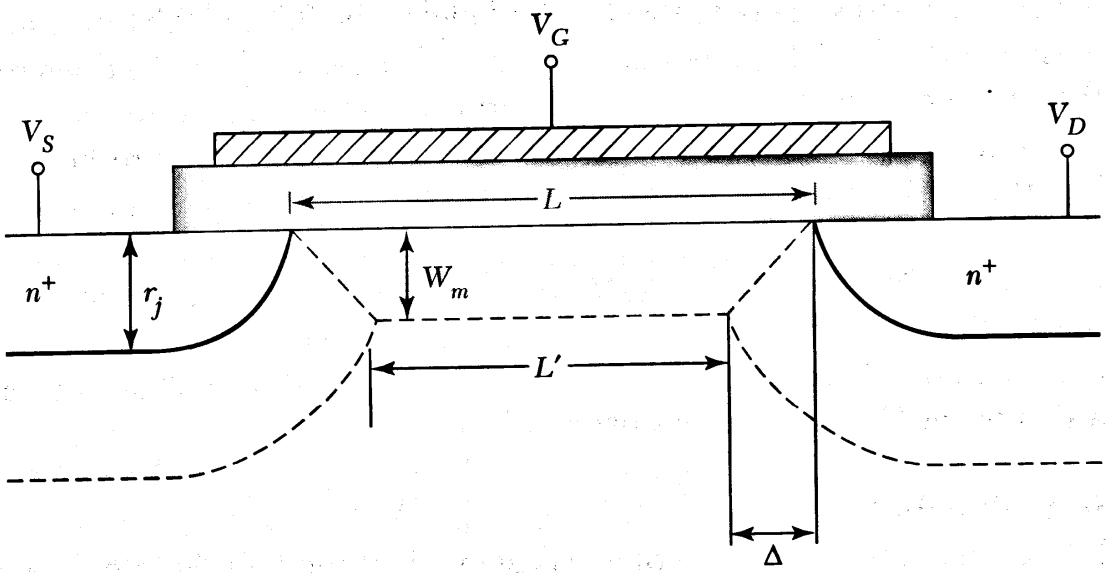
Tali sovrapposizioni risultano trascurabili nel caso di MOS "lunghi" (di vecchia generazione), ma diventano importanti nel caso di MOS "corti".

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"

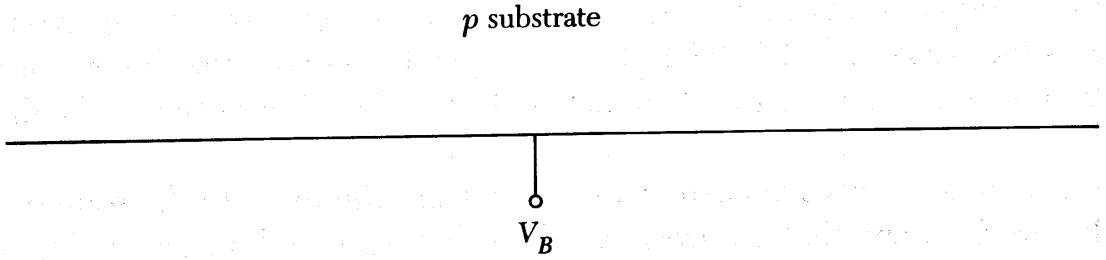
(il che equivale a dire che **la tensione di soglia V_{TH} si riduce al diminuire di L**).

Effetti di canale corto

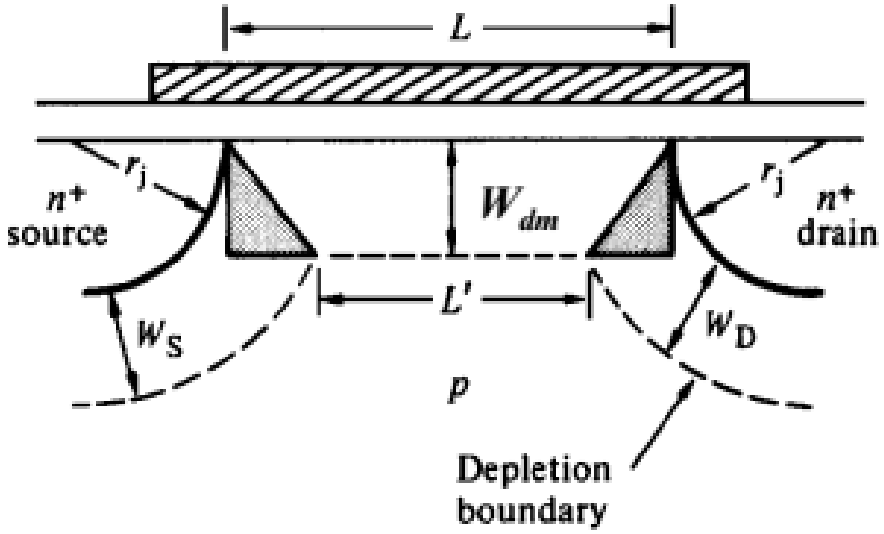




Quanta carica del canale viene controlla dal gate?



Variazione della tensione di soglia (Vt roll-off)



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

La carica risulta ridotta di un fattore

$$1-\frac{L+L'}{2L}$$

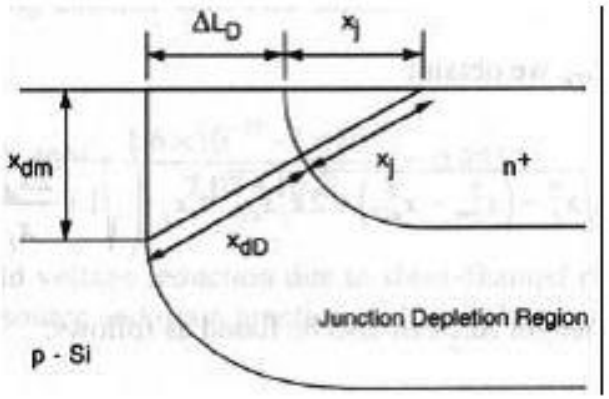
Variazione della tensione di soglia (Vt roll-off)

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

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

$$x_{dS} = \sqrt{\left(\frac{2\varepsilon_{Si}}{qN_A}\right)(\phi_0)}$$

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



Si trova che:

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

Risolvendo otteniamo:

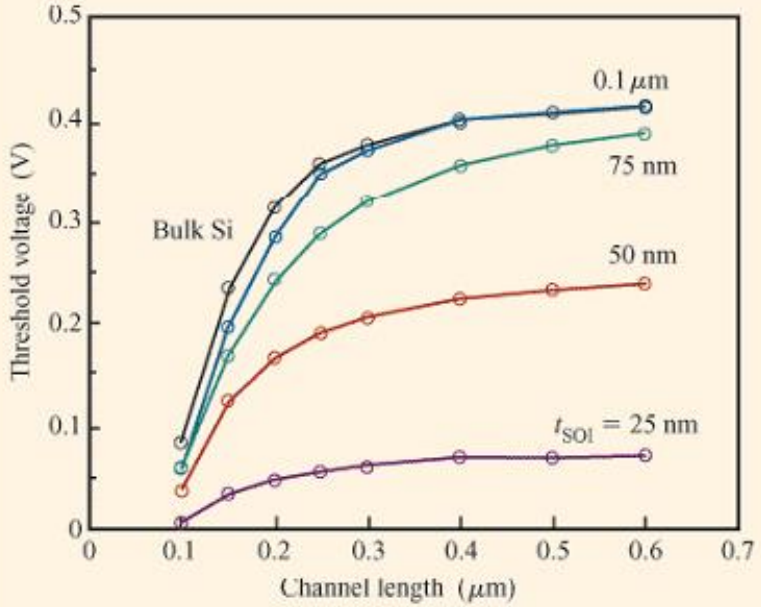
$$\Delta L_S \cong x_j\left(\sqrt{1 + \frac{2x_{dS}}{x_j}} - 1\right)$$

Variazione della tensione di soglia (Vt roll-off)

Ne segue che la variazione della tensione di soglia sarà:

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

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

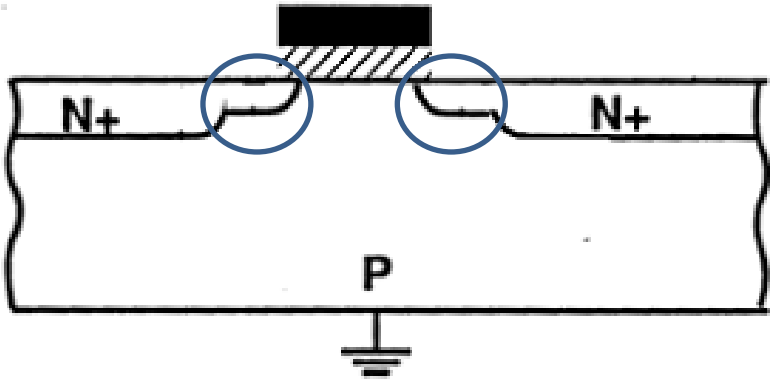


Variazione della tensione di soglia (Vt roll-off)

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

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!

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

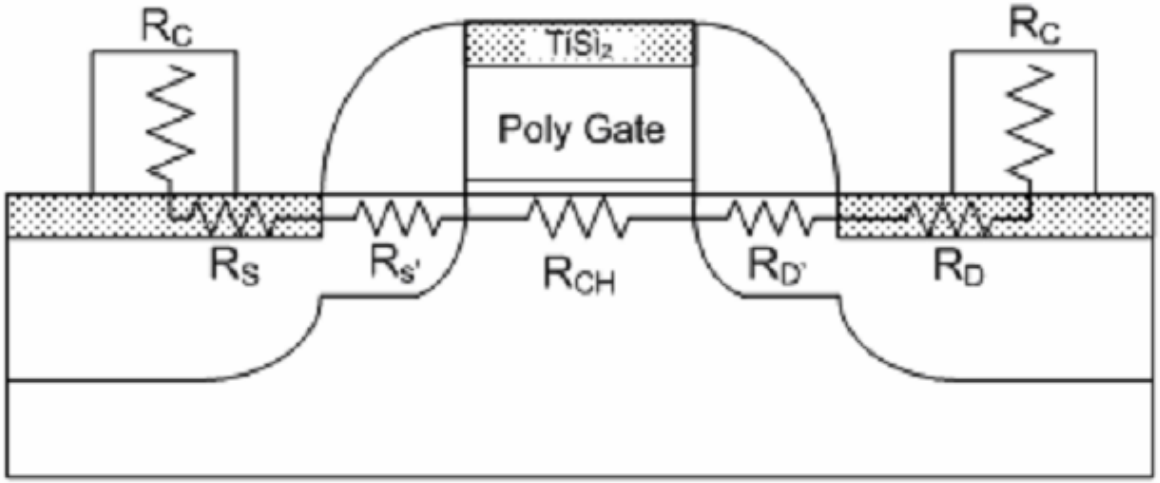
Variazione della tensione di soglia (Vt roll-off)

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

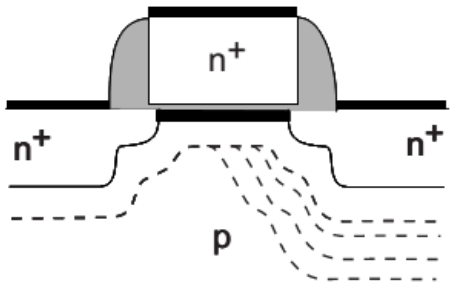


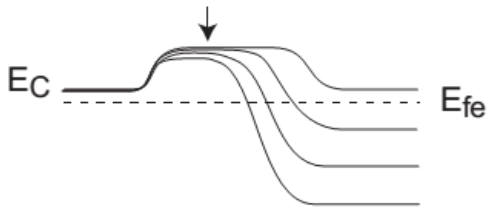
Drain-Induced Barrier Lowering

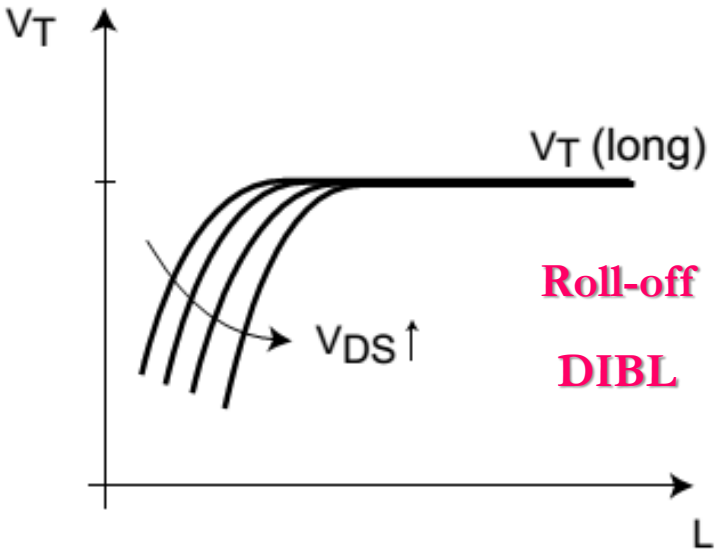
Quando siamo in condizioni di svuotamento, o al massimo in debole inversione, all'aumentare della tensione di gate aumenta la regione di svuotamento al drain, e **se il canale è corto è possibile che source e drain si accoppino**

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

Aumento **significativo della corrente di sottosoglia**

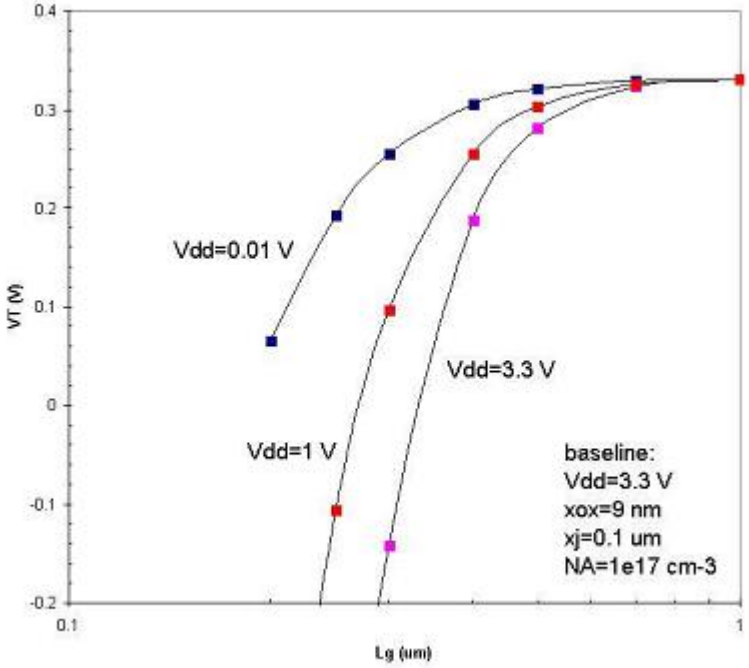


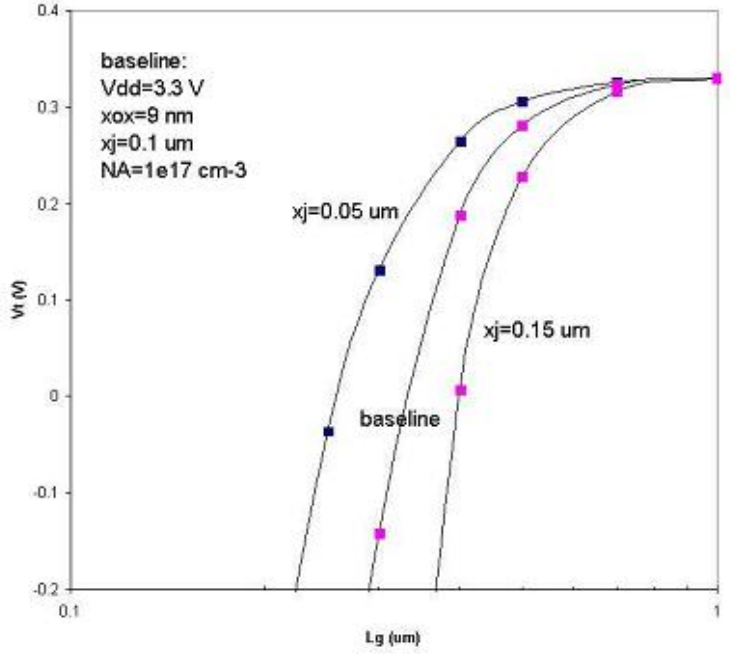




Drain-Induced Barrier Lowering

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.**

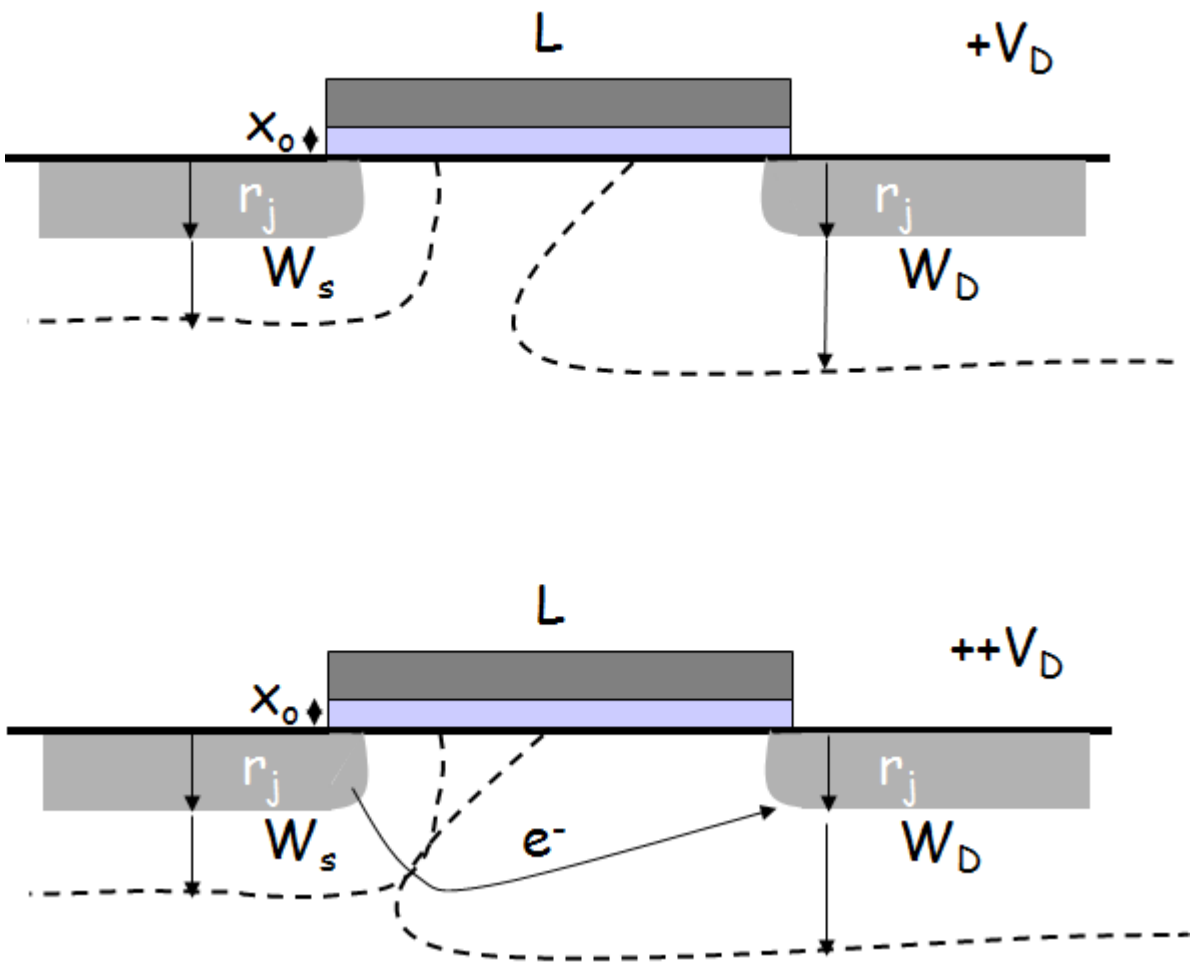
È chiaro allora che, con il procedere della miniaturizzazione, **la possibilità della "perforazione diretta"** (regioni di svuotamento D-B e S-B che si congiungono) **prende ad aumentare.**

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

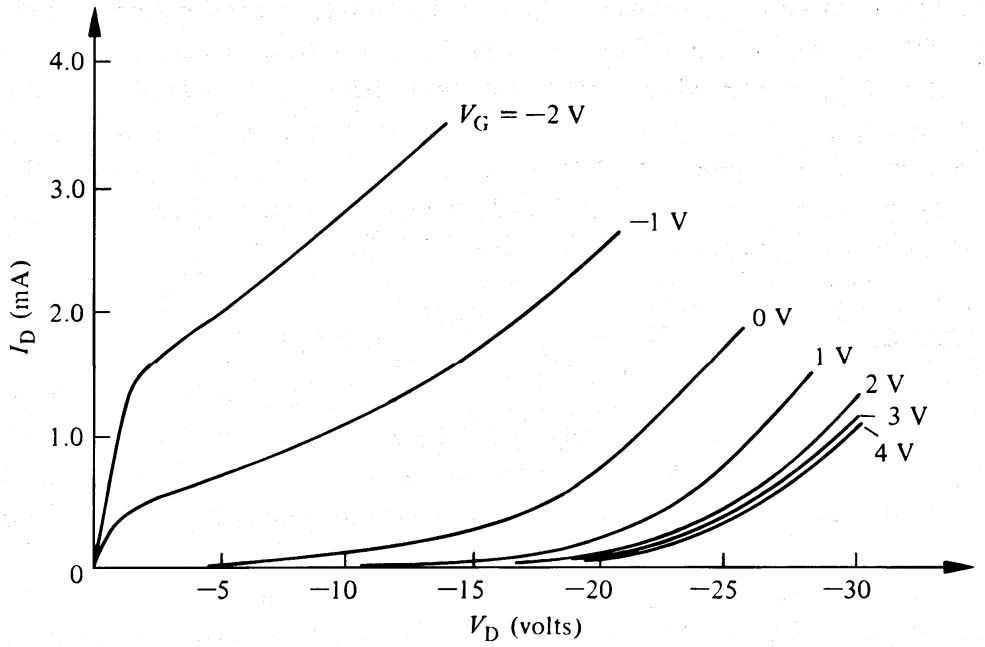
$$d = \sqrt{\frac{2\varepsilon_S \left(V_j + V_R\right)}{qN_A}}$$

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!

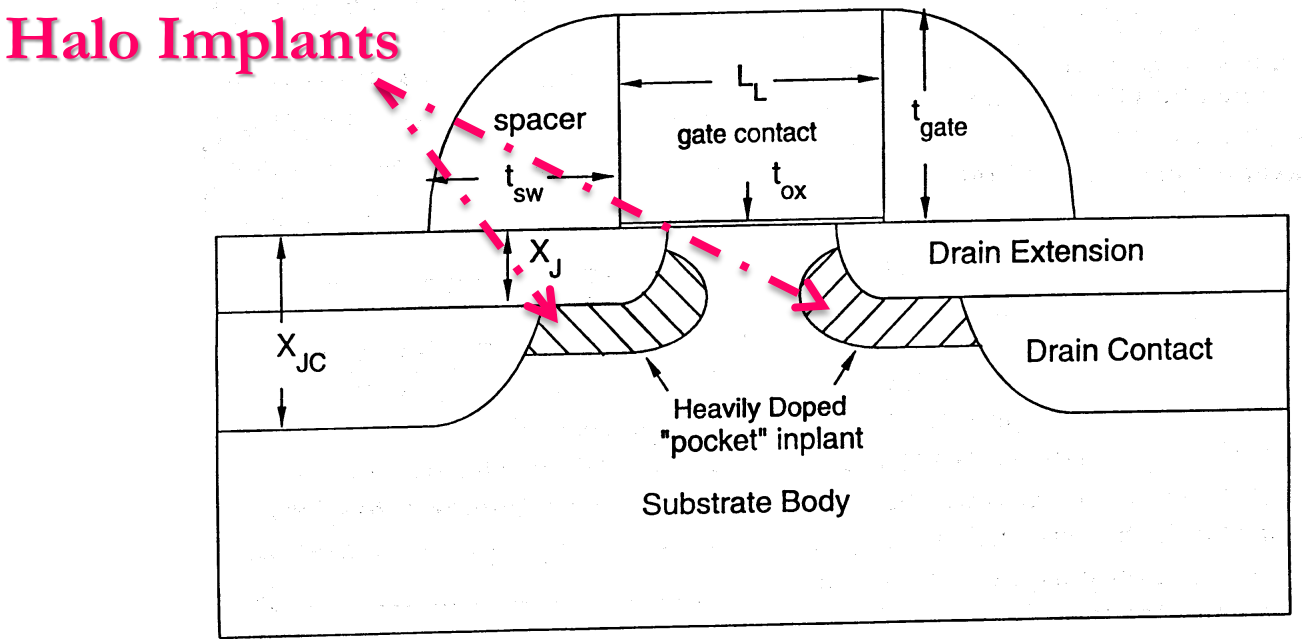
La tensione di polarizzazione scala infatti di un fattore S^2

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

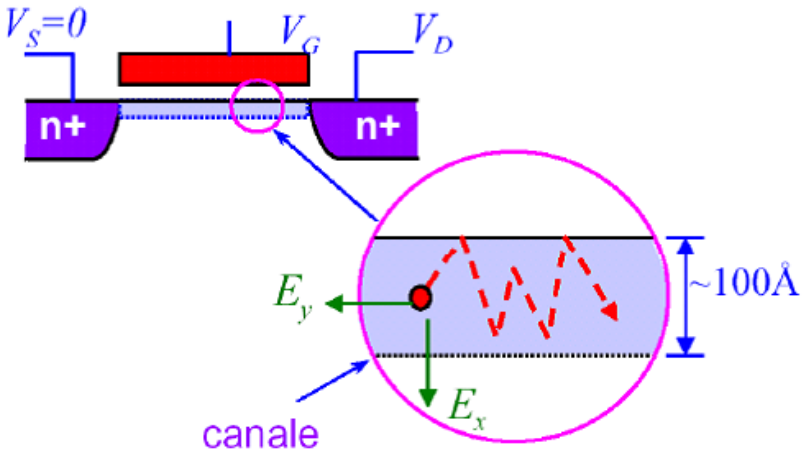
Degradazione della mobilità dovuta alle collisioni superficiali

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 μph:** 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

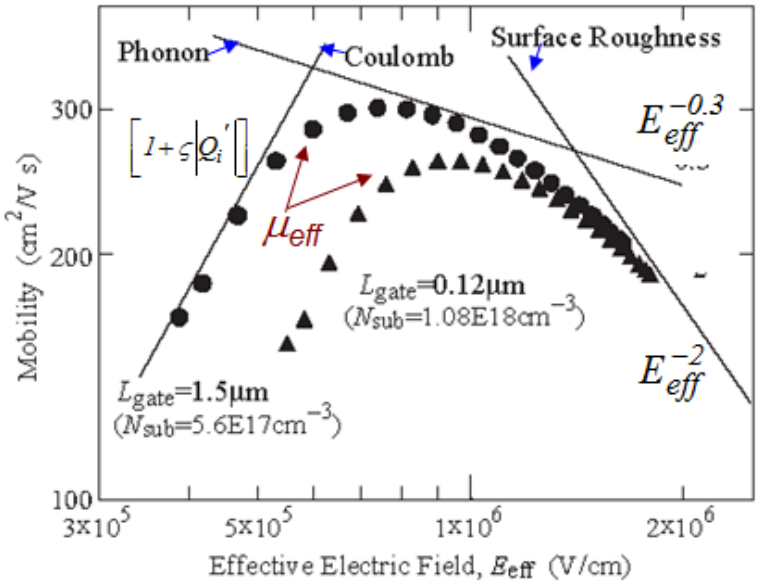
Anche per il campo longitudinale (E_X) la mobilità è descritta da tre contributi

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

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



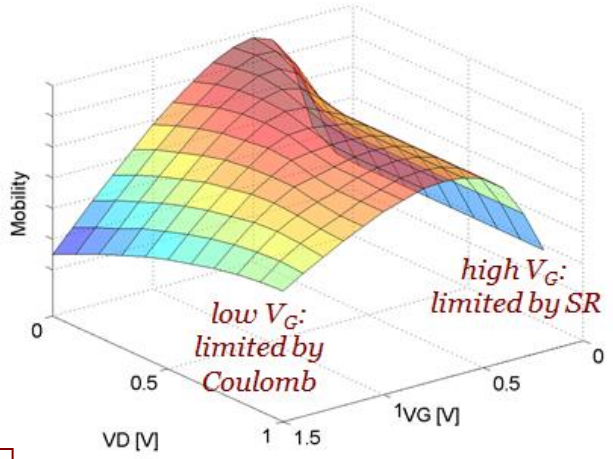
Ad eccezione del primo termine, **in generale la mobilità diminuisce all'aumentare del campo**

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

Velocity saturation

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**.

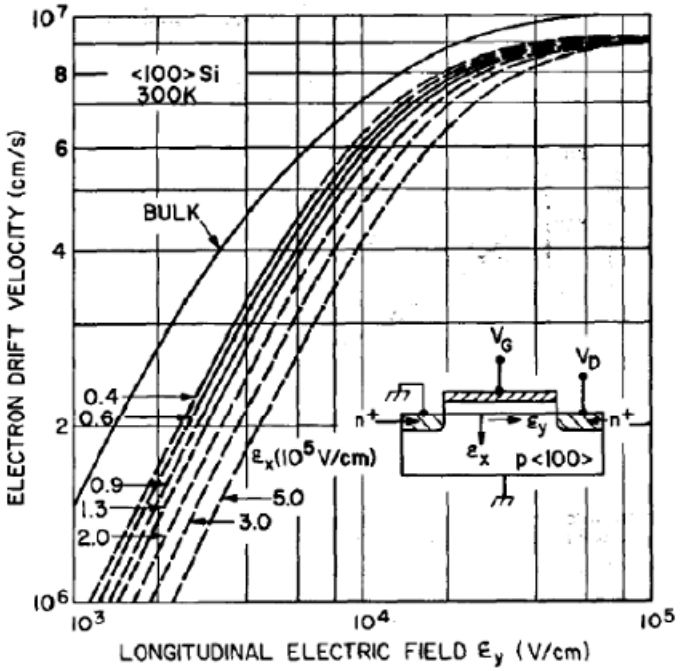
Effetti di canale corto

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



Effetti di canale corto

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

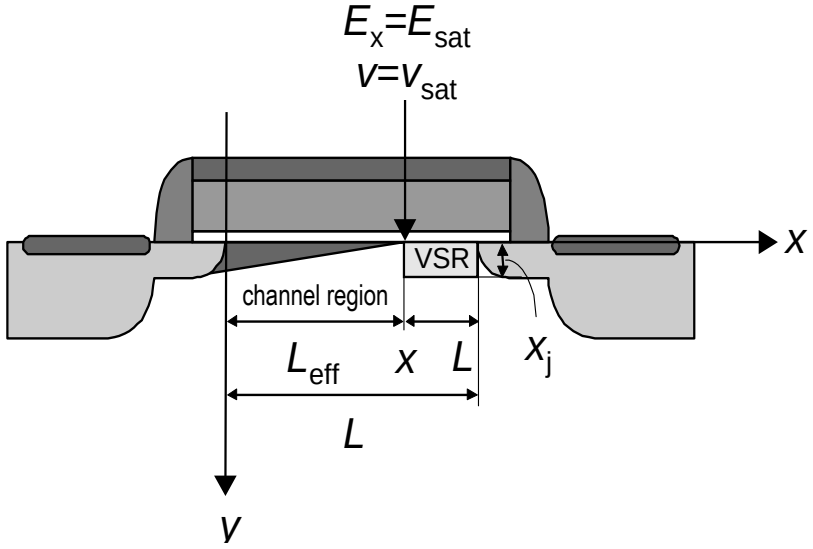
Modulazione della lunghezza di canale

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

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

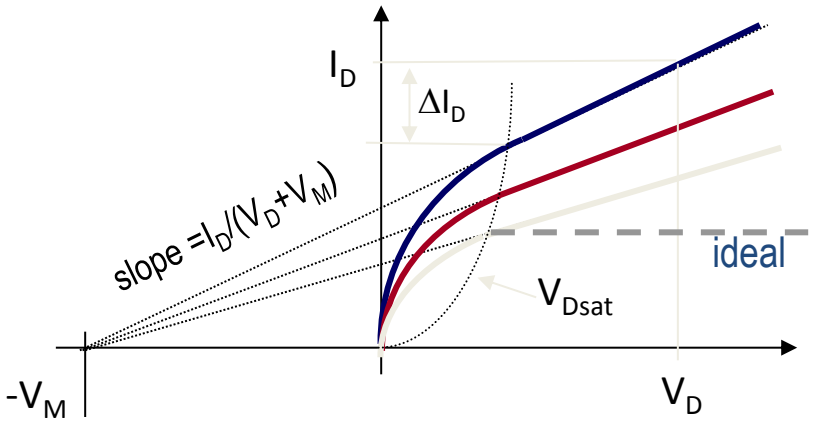
$$L_C = \sqrt{\frac{\varepsilon_{si} \cdot X_J}{C_{OX}'}}$$

junction depth

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

La conduttanza di canale diventa

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



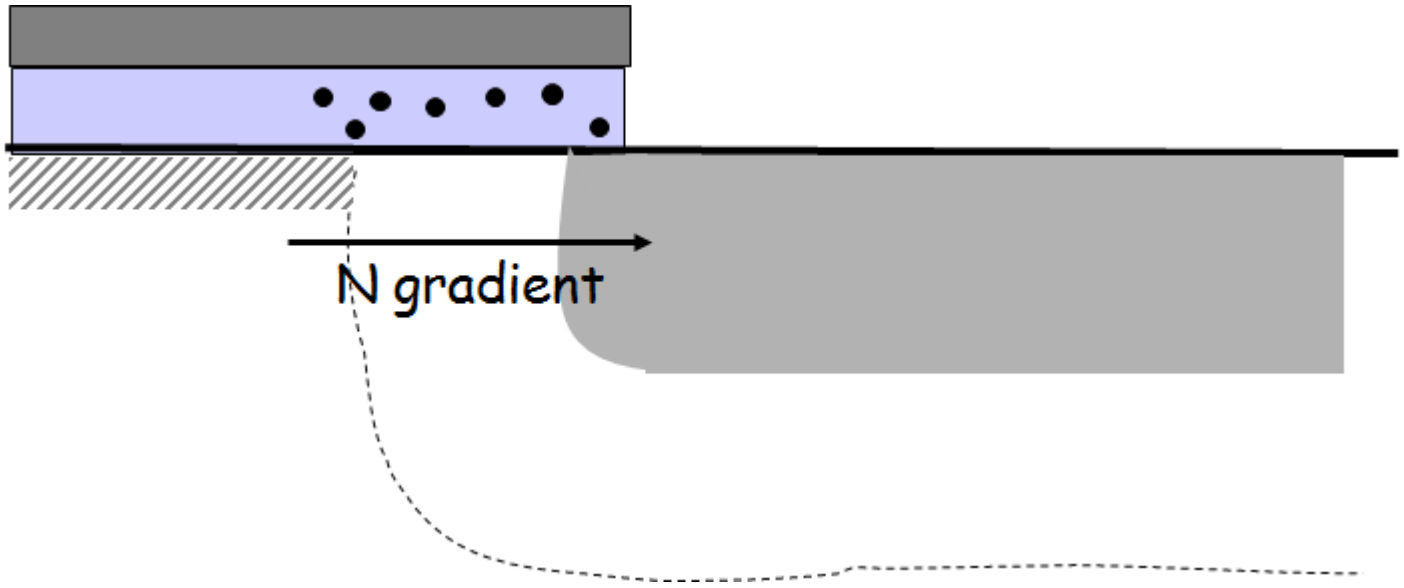
$$\frac{\Delta L}{L} = \lambda V_{DS}$$

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

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:

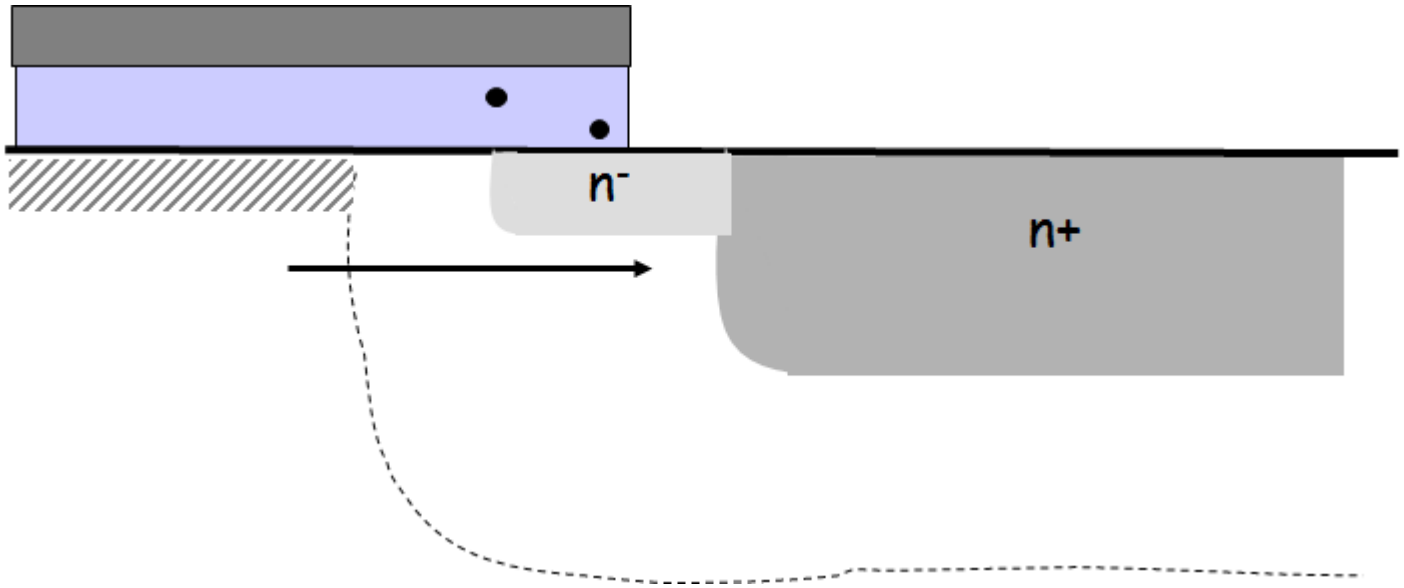
- una **variazione incontrollata della V_{TH}**, che dipende dalla carica nell'ossido attraverso la "tensione di banda piatta" V_{FB};
- una **possibile rottura "a lungo termine" dell'ossido**, la cui qualità degrada nel tempo all'aumentare dell'iniezione di elettroni caldi.

  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